2017

Essential Genes And Their Role In Autism Spectrum Disorder

Xiao Ji
University of Pennsylvania, jixiao@mail.med.upenn.edu

Essential Genes And Their Role In Autism Spectrum Disorder

Abstract
Essential genes (EGs) play central roles in fundamental cellular processes and are required for the survival of an organism. EGs are enriched for human disease genes and are under strong purifying selection. This intolerance to deleterious mutations, commonly observed haploinsufficiency and the importance of EGs in pre- and postnatal development suggests a possible cumulative effect of deleterious variants in EGs on complex neurodevelopmental disorders. Autism spectrum disorder (ASD) is a heterogeneous, highly heritable neurodevelopmental syndrome characterized by impaired social interaction, communication and repetitive behavior. More and more genetic evidence points to a polygenic model of ASD and it is estimated that hundreds of genes contribute to ASD. The central question addressed in this dissertation is whether genes with a strong effect on survival and fitness (i.e. EGs) play a specific role in ASD risk. I compiled a comprehensive catalog of 3,915 mammalian EGs by combining human orthologs of lethal genes in knockout mice and genes responsible for cell-based essentiality. With an updated set of EGs, I characterized the genetic and functional properties of EGs and demonstrated the association between EGs and human diseases. Next I provided evidence for a stronger contribution of EGs to ASD risk, compared to non-essential genes (NEGs). By examining the exonic de novo and inherited variants from 1,781 ASD quartet families, I demonstrated a significantly higher burden of damaging mutations in EGs in ASD probands compared to their non-ASD siblings. Analysis of EGs in the developing brain identified clusters of co-expressed EGs implicated in ASD, among which I proposed a priority list of 29 EGs with potential ASD risk as targets for future functional and behavioral studies. Finally, I developed the essentiality burden score (EBS), which captures the burden of rare mutations in EGs as a novel polygenic predictor of individual ASD risk and a useful addition to the current tools for understanding the polygenic architecture of ASD. Overall, I show that large-scale studies of gene function in model organisms and human cell lines provide a powerful approach for prioritization of genes and pathogenic variants identified by sequencing studies of complex human disease.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Maja Bucan

Keywords
Autism spectrum disorder, Essential genes, Mouse knockout, Mutational burden

Subject Categories
Bioinformatics | Genetics

ESSENTIAL GENES AND THEIR ROLE IN AUTISM SPECTRUM DISORDER

Xiao Ji

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Maja Bucan, Ph.D., Professor of Genetics

Graduate Group Chairperson

_____

Li-San Wang, Ph.D., Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee

Hakon Hakonarson, M.D., Ph.D., Professor of Pediatrics

Christopher D. Brown, Ph.D., Assistant Professor of Genetics

Nancy Zhang, Ph.D., Associate Professor of Statistics

Santhosh Girirajan, Ph.D., Assistant Professor of Biochemistry & Molecular Biology and Assistant Professor of Anthropology

ESSENTIAL GENES AND THEIR ROLE IN AUTISM SPECTRUM DISORDER

ACKNOWLEDGMENT

It would not have been possible to complete this dissertation without the support that I received from many people.

I would like to express my appreciation and thanks to my advisor Dr. Maja Bucan. Since I joined her lab six years ago as an inexperienced student looking for initial research experiences, she began to take me hand in hand through everything that I needed to learn to be a real scientist. She is always encouraging and supportive during every step I took. It is fortunate for me to have her as my dissertation advisor and mentor.

I would like to thank my thesis committee members, Dr. Hakon Hakonarson, Dr. Christopher Brown, Dr. Nancy Zhang and Dr. Santhosh Girirajan for devoting their time for our regular meetings and providing invaluable guidance to the development of this dissertation. Special thanks go to Dr. Christopher Brown for our one-on-one seminar on the history of genome-wide association study, from which I gained in-depth perspective on my dissertation proposal. I would also like to thank Dr. Rachel Kember. During the years in the lab, she was like a co-mentor to me and she was always there whenever I needed help. In addition to all of our discussions in research, she was an excellent English teacher of me.

I am grateful to my parents, who encouraged me to pursue a Ph.D. degree and offered me their support without restraint. Finally I thank my wife, Dr. Lu Chen, who has been always together with me through the ups and downs during these years. She also provided important insights into a number of statistical problems in the dissertation.

ABSTRACT

ESSENTIAL GENES AND THEIR ROLE IN AUTISM SPECTRUM DISORDER

Xiao Ji

Maja Bucan

Essential genes (EGs) play central roles in fundamental cellular processes and are required for the survival of an organism. EGs are enriched for human disease genes and are under strong purifying selection. This intolerance to deleterious mutations, commonly observed haploinsufficiency and the importance of EGs in pre- and postnatal development suggests a possible cumulative effect of deleterious variants in EGs on complex neurodevelopmental disorders. Autism spectrum disorder (ASD) is a heterogeneous, highly heritable neurodevelopmental syndrome characterized by impaired social interaction, communication and repetitive behavior. More and more genetic evidence points to a polygenic model of ASD and it is estimated that hundreds of genes contribute to ASD. The central question addressed in this dissertation is whether genes with a strong effect on survival and fitness (i.e. EGs) play a specific role in ASD risk. I compiled a comprehensive catalog of 3,915 mammalian EGs by combining human orthologs of lethal genes in knockout mice and genes responsible for cell-based essentiality. With an updated set of EGs, I characterized the genetic and functional properties of EGs and demonstrated the association between EGs and human diseases. Next I provided evidence for a stronger contribution of EGs to ASD risk, compared to non-essential genes (NEGs). By examining the exonic *de novo* and inherited variants from 1,781 ASD quartet families, I demonstrated a significantly higher burden of

damaging mutations in EGs in ASD probands compared to their non-ASD siblings. Analysis of EGs in the developing brain identified clusters of co-expressed EGs implicated in ASD, among which I proposed a priority list of 29 EGs with potential ASD risk as targets for future functional and behavioral studies. Finally, I developed the essentiality burden score (EBS), which captures the burden of rare mutations in EGs as a novel polygenic predictor of individual ASD risk and a useful addition to the current tools for understanding the polygenic architecture of ASD. Overall, I show that large-scale studies of gene function in model organisms and human cell lines provide a powerful approach for prioritization of genes and pathogenic variants identified by sequencing studies of complex human disease.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: Introduction

One of the central goals of human genetics studies is to understand the genetic contribution to human diseases. This knowledge is of great value in combating disease and promoting human health. The genes responsible for a wide range of Mendelian disorders, such as sickle cell anemia, cystic fibrosis, and Huntington's disease, have been well understood by the genetics community (Stenson et al., 2014). However, it has been a challenge to identify risk genes and variants underlying a majority of common complex diseases including cardiovascular diseases, cancer, diabetes and psychiatric disorders, where complicated interactions of multiple genes and environmental factors are involved in their etiology (Risch, 2000; Wray et al., 2014). There has been an ongoing debate in the field of genetics over how genetic variations contribute to the risk of common complex diseases. The 'common disease-common variant' hypothesis predicts that common genetic variants with low penetrance are the major contributors of individual susceptibility to complex diseases. In contrast, the 'common disease-rare variant' hypothesis argues that the risk of complex disease is mainly due to rare variants that are more specific to individuals with relatively high penetrance. Both hypotheses have their place in current research and are supported by substantial evidence, therefore it is important to evaluate the contribution of both common and rare variants to the risk of complex diseases (Gibson, 2012; Schork et al., 2009). The key to identifying genetic variants contributing to complex diseases is to pinpoint risk genes and variants from a huge number of those that are biologically insignificant or irrelevant for that disease.

Important advances in the study of complex disease were the development of technologies that enable systematic interrogation of many genetic variants in large cohorts of patients. Single nucleotide polymorphism (SNP) array-based genome-wide association studies (GWAS) represent a powerful tool for uncovering the common genetic variants that underlie risk of complex diseases. Furthermore, next generation sequencing technologies enable the investigation of the role of low-frequency or rare variants in complex diseases, which may explain additional disease risk or trait variability. However, despite the genetic associations discovered though the studies of both common and rare variants, a full understanding of the genetic architecture of most complex disorders has yet to be achieved.

One of the substantial challenges for current sequencing-based association studies of complex diseases comes from the limitation of the classical single variant-based association test, where limited sample sizes, modest effect sizes of variants and the multiple testing burden restricts its statistical power (Lee et al., 2014). As an alternative approach, aggregation tests that evaluate the cumulative effect of multiple variants in a gene or region can increase statistical power when a group of variants are associated with a disease or trait of interest (Lee et al., 2014; Li and Leal, 2008; Madsen and Browning, 2009; Wu et al., 2011). Following the concept of gene- or region-level aggregation tests, a top-down strategy starts from identifying a large set of genes with key characteristics that are known to play a role in studied diseases. Groups of variants in these candidate gene sets are then jointly tested in order to increase statistical power. This strategy was applied by a number of recent genetic studies of schizophrenia. For example, Purcell et

al. performed a polygenic burden test of rare disruptive mutations in schizophrenia candidate gene sets (including synaptic genes, voltage-gated calcium channel genes and targets of the fragile X mental retardation protein) using exome sequences of ~2,500 schizophrenia cases and ~2,500 controls (Purcell et al., 2014). In addition, copy number variant (CNV) burden within gene sets involved in neurodevelopmental or neurological function was assessed in a schizophrenia cohort of ~20,000 cases and ~20,000 controls (Cnv et al., 2017). These studies provided a proof-of-principle that the candidate gene set approach which evaluates the aggregational effect of multiple variants can facilitate the discovery of risk alleles in neuropsychiatric diseases.

Of all of the genes in the genome, there is a subset of essential genes (EGs) that play central functional roles and are required for the survival of an organism. The identification and characterization of the core set of genes that are necessary for basic developmental functions, i.e. the “essentialome”, is an important biological question by itself, as it provides insights into the molecular basis for key biological processes in multiple organisms including human (Zhan and Boutros, 2016). In *S. cerevisiae* (budding yeast), ~ 20% of ~ 6,000 genes are necessary for viability and proliferation in rich medium (Giaever et al., 2002; Winzeler et al., 1999). However, in addition to the core set of EGs that result in lethal phenotype upon loss, there are other genes that are conditionally essential. These genes have been extensively studied in *S. cerevisiae*. It has been shown that yeast mutants with one of these genes deleted are sensitive to additional perturbations such as stress conditions (Giaever et al., 2002), chemicals (Costanzo et al., 2010; Hillenmeyer et al., 2008) and knock out of a second gene (i.e. synthetic lethality)

(Costanzo et al., 2010; Nijman, 2011). Therefore, deciding whether these genes also count as essential is a matter for discussion. In multicellular organisms, many EGs are housekeeping genes that are required for maintaining basal cellular functions and tend to be ubiquitously expressed at constant levels in all cell types (Eisenberg and Levanon, 2013). However, some other EGs in multicellular organisms can be restricted to the function of specific tissues or certain developmental stages (Zhan and Boutros, 2016). For example, mice with targeted disruption of *Fatp4* gene that encodes a fatty acid transport protein died shortly after birth because of a skin defect (Herrmann et al., 2003). The *Fatp4* knockout mice could be rescued by introducing transgenic expression of *Fatp4* in skin cells (Shim et al., 2009). In contrast, adipocyte-specific inactivation of *Fatp4* did not result in severe phenotypes in mice (Lenz et al., 2011), showing that *Fatp4* is likely to be required for the proper function of a single tissue, whereas it is essential for the viability of the whole organism. In the scope of this dissertation, I defined an EG as a gene that causes lethality of a multicellular organism when fully knocked-out, whether the gene is essential in all tissues or not.

Historically, forward genetics strategies based on chemically induced, radiation-induced or insertional mutagenesis had been extensively applied to investigate the link between genotypes and phenotypes (Zhan and Boutros, 2016). In *C. elegans*, Clark et al. and Johnsen & Baillie independently identified hundreds of lethal mutations in specific chromosomal regions using ethyl methanesulfonate (EMS) mutagenesis. They estimated that the total number of EGs in *C. elegans* is at least 2,850~3,500, which accounts for 15~18% of all protein coding genes in the *C. elegans* genome (Clark et al., 1988; Johnsen

and Baillie, 1991). In *D. melanogaster*, P-transposable element has been widely used to disrupt Drosophila genes. For instance, the Berkeley *Drosophila* Genome Project generated mutant lines for 40% of *Drosophila* genes and observed that 8~16% of genes led to lethal phenotypes when disrupted (Bellen et al., 2004).

Built on the foundation of the completed genome sequences of many model organisms, reverse genetics approaches enabled the exploration of gene essentiality on a genome-wide scale. RNA interference (RNAi) based gene silencing has been proved to be a successful strategy in discovering EGs. Kamath et al. used RNAi to inhibit the function of ~16,700 genes in *C. elegans*, and identified mutant phenotypes of 1,722 genes. 68% of these genes (n=1,170) exhibited nonviable RNAi phenotypes (Kamath et al., 2003). Boutros et al. performed genome-wide RNAi analysis of the growth and viability in *Drosophila* cells and identified 438 EGs, among which 80% lacked known mutant alleles in *Drosophila* (Boutros et al., 2004). Dietzl et al. generated a genome-wide library of 22,270 RNAi transgenic *Drosophila* lines that covered 88% of the predicted protein-coding genes in *Drosophila*, among which 17.5% of transgenic lines exhibited lethal phenotypes (Dietzl et al., 2007). While these studies generated sizable catalogs of EGs in studied organisms, the percentages of EGs discovered varied in these studies due to common limitations of RNAi screens such as variability and incompleteness of knockdowns as well as potential nonspecificity of RNAi targets (Boutros and Ahringer, 2008). Regardless, these studies do not necessarily contradict the estimation from Miklos & Rubin that around one third of genes are essential for viability in these model organisms (Miklos and Rubin, 1996).

Identification of EGs in the mouse is of particular interest due to the evolutionary closeness between mouse and human, as well as the great potential of mouse models in translational research. Over decades, the genetics community collected a substantial amount of phenotypic data in knockout mouse strains generated by both forward and reserve genetics approaches such as ethyl-nitrosourea (ENU) mutagenesis, transposon mutagenesis, gene trapping and gene targeting in mouse embryonic stem cells (Eppig et al., 2005). Based on reported homozygous embryonic/perinatal lethal mouse mutants, it was estimated that ~30% of mouse genes are essential for mouse viability (Dickinson et al., 2016; White et al., 2013). Due to the extensive similarity between the genomes of mouse and human, human EGs can be inferred from the human orthologs of prenatal or preweaning lethal genes in the mouse (Dickerson et al., 2011; Feldman et al., 2008; Georgi et al., 2013; Goh et al., 2007; Park et al., 2008). We are particularly interested in the potential connection between EGs and human disease. Earlier studies of human orthologs of EGs in the mouse proposed that the majority of human disease genes are non-essential, because mutations in EGs prevent viability and thus do not contribute to human disease (Domazet-Loso and Tautz, 2008; Feldman et al., 2008; Goh et al., 2007; Park et al., 2008) . However, the role of EGs in human disease could be underestimated, since some of these studies also presented contrasting evidence showing that human disease genes can also display some characteristics of EGs, such as high connectivity in gene networks (Goh et al., 2007) and an early evolutionary emergence (Domazet-Loso and Tautz, 2008). More recent studies on human orthologs of EGs in the mouse began to redefine the role of EGs in human disease. With an analysis of the overlap between 1,299 EGs and known human disease genes, Dickerson et al. pointed out that EGs actually

comprised a major portion of disease genes (Dickerson et al., 2011). Georgi et al. reinforced this notion by showing an enrichment of disease genes among an updated list of 2,472 human orthologs of EGs in the mouse (Georgi et al., 2013).

To better understand to role of EGs in human disease, it is helpful to clarify the difference in mutational spectrums of EGs and non-essential genes (NEGs). In a disease-associated NEG, we may observe disease phenotypes when homozygous loss-of-function mutations or compound heterozygosity of null alleles occur in an individual. However in an EG, we won't observe homozygous loss-of-function mutations in living individuals because they cause lethality. Instead, EGs could contribute to human disease through milder alleles other than functionally null alleles (Figure 1.1). It has been shown that EGs exhibit a reduced number of exonic missense (Georgi et al., 2013; Petrovski et al., 2013) and loss-of-function (Lek et al., 2016) variants in general population, as well as a shift in allele frequency towards rare alleles (Georgi et al., 2013). Moreover, previous studies showed evidence that EGs are prone to exhibiting haploinsufficiency (Deutschbauer et al., 2005; Georgi et al., 2013), which suggests that heterozygous alleles in EGs are more likely to be deleterious and pathogenic. These observations support the functional importance of EGs in humans and implicate that EGs are more likely to have functional consequences when mutated.

In this dissertation, I investigated the connection between EGs and human diseases, with a focus on a neurodevelopmental disease - autism spectrum disorder (ASD). ASD is a neurodevelopmental disorder characterized by repetitive behavior and impairments in social interaction, communication and language (2013). The signs of autism begin to

appear over the first year of life (Ozonoff et al., 2008). According to the latest survey from the Centers for Disease Control and Prevention, the prevalence of ASD is 1 in 68, and males are 4.5 times more likely to develop ASD compared to females (Christensen, 2016). There is general agreement across family and twin studies that the heritability of ASD is between ~60% to ~90% (Bailey et al., 1995; Folstein and Rutter, 1977; Hallmayer et al., 2011; Lichtenstein et al., 2010; Ronald and Hoekstra, 2011; Sandin et al., 2014). The genetic causes of ASD are highly heterogeneous among patients, and identified ASD linked mutations accounting for more than 1% of ASD cases are very rare (Jeste and Geschwind, 2014; State and Sestan, 2012). It has been demonstrated that common variants carry a substantial ASD risk (Anney et al., 2012; Gauglerl et al., 2014; Klei et al., 2012; Wang et al., 2009), which supports the 'common disease-common variant' hypothesis in ASD. Based on this hypothesis, a number of ASD risk loci were discovered through genetic linkage analysis (Szatmari et al., 2007; Weiss et al., 2009) and GWAS (Anney et al., 2010; Ma et al., 2009; Wang et al., 2009). However, it is difficult to identify and replicate the ASD-linked common variants by these traditional genetic tests due to their small effect sizes and currently limited sample size (~5,000 ASD cases). Based on the "common disease-rare variant" hypothesis, many ASD studies that focused on protein-disrupting, rare *de novo* variants in affected children have successfully implicated hundreds of single nucleotide variants (SNVs) and copy number variants (CNVs) as potential ASD risk factors (Bucan et al., 2009; De Rubeis et al., 2014; Gilman et al., 2011; Glessner et al., 2009; Gratten et al., 2013; Griswold et al., 2012; Iossifov et al., 2014; Iossifov et al., 2012; Itsara et al., 2010; Levy et al., 2011; Marshall et al., 2008; Neale et al., 2012; O'Roak et al., 2012a; O'Roak et al., 2012b; Pinto et al.,

2010a; Sanders et al., 2011; Sanders et al., 2012; Sebat et al., 2007; Szatmari et al., 2007). However, a large proportion of ASD heritability remains unexplained and the genetic mechanisms involved in ASD are still not fully understood. As a way to explain the complexity of the genetic architecture of ASD, more and more genetic evidence points to a polygenic model of ASD (de la Torre-Ubieta et al., 2016), i.e. at least hundreds of genes and a large number of common variants with modest effect, and rare or *de novo* variants with strong effect, contributing to ASD risk collaboratively.

The early on-set of ASD suggests a prenatal or early postnatal origin. Multiple lines of evidence implicated that impairments of early brain development were involved in the pathogenesis of ASD (Parikshak et al., 2013; Stoner et al., 2014; Willsey et al., 2013a). For instance, Parikshak et al. found that ASD genes from multiple sources converged on pathways implicated in prenatal and early post natal synaptic development (Parikshak et al., 2013). Willsey et al. reported convergence of ASD genes on midfetal deep cortical projection neurons (Willsey et al., 2013a). Stoner et al. observed disorganization of neurons in prefrontal and temporal cortical tissues in 10 of 11 autistic children and suggested that such abnormality emerged at prenatal developmental stages (Stoner et al., 2014). Therefore, EGs as a group of genes that are required for normal pre- and postnatal development are prime candidates for the analysis of the polygenic architectures of ASD.

The objective of this dissertation is to systematically investigate the potential link between EGs and ASD, which was proposed by Georgi et al., who observed that genes with *de novo* events in ASD patients are enriched for EGs (Georgi et al., 2013). Furthermore, the two hit model of neuropsychiatric disorders was initially proposed by

Girirajan & Eichler to explain the phenotypic variability among patients. This model suggests that in a network of genes in a pathway associated with neuropsychiatric disorders, a single hit initially disrupts the pathway and results in a milder phenotype, and a second hit further damages the pathway to generate a much more severe phenotype (Girirajan and Eichler, 2010). Therefore, in neurodevelopmental disorders where individual candidate genes cannot fully explain their genetic basis, it is possible that multiple deleterious variants in EGs constitute a genetic background that influences an individual's disease risk. I hypothesized that a cumulative effect of a range of alleles in EGs may contribute to developmental or behavior anomalies such as ASD. In this thesis, my aims are to address these challenges as follows.

In **Chapter 2**, I identified and compiled the most comprehensive set of EGs to-date by combining data from cell-based assays in human cell lines and systematically phenotyped knock-out mice. I characterized the genetic and functional properties of EGs and demonstrated the association between EGs and human disease.

In **Chapter 3**, I provided compelling evidence for a significant contribution of EGs to ASD risk compared to NEGs by showing a higher burden of damaging mutations in EGs in ASD probands and enrichment of EGs among currently known ASD risk genes. Moreover, I identified clusters of co-expressed EGs implicated in ASD through the analysis of EGs in the developing brain.

In **Chapter 4**, I developed the essentiality burden score (EBS), based on exonic rare variants in EGs, as a novel predictor to ASD risk. I compared EBS, polygenic risk score

(PRS), and rare CNV burden to evaluate their performance in predicting ASD disease risk. Furthermore, I investigated the interplay between EBS and rare variants in a high-penetrant ASD risk gene, *NRXN1*.

Figures

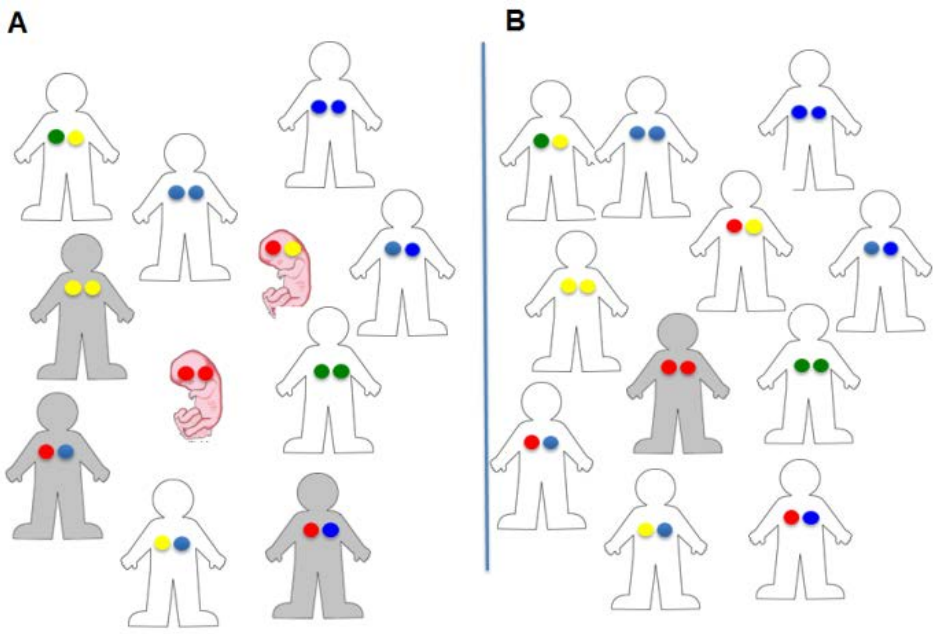

Figure 1.1 Distinct mutational spectrums of variants in essential and non-essential genes. Loss-of-function variants (in red) in both alleles (circles) of essential genes **(A)** lead to lethality or miscarriages when homozygous (depicted as embryo) and are likely to lead to a disease phenotype (shaded gray) when they are heterozygous. Loss-of-function variants in many non-essential genes **(B)** produce a disease phenotype when homozygous and no disease phenotype when heterozygous (shaded white). Hypomorph alleles (yellow) in essential genes may produce a disease phenotype when homozygous or may lead to lethality when combined with another hypomorph or loss-of-function allele in the same gene. Benign alleles are depicted with different colored circles.

CHAPTER 2: Characterization of essential genes

Introduction

Research of gene essentiality has potential implications for the genetic basis of human disease. However, it is difficult to identify EGs directly from human studies because null mutations in EGs are missing in living individuals. A number of EGs were implicated through case studies of families with mutations in genes linked to miscarriages or lethal birth defects (Malfatti et al., 2014; Michalk et al., 2008; Stangenberg et al., 1992). Because of low sample sizes available for these case studies, the number of human EGs discovered through this method is limited. For example, Stangenberg et al. reported a patient with recurrent miscarriages who delivered a hydropic stillborn infant with β-Glucuronidase (*GUSB*) deficiency (Stangenberg et al., 1992). Michalk et al. found that one fetus had homozygous loss-of-function mutations in *CHRNA1* which could disable the function of acetylcholine receptor and lead to intrauterine death (Michalk et al., 2008). Moreover, Malfatti et al. reported five *NEB*-mutated infants who presented severe congenital myopathy leading to death in the first day after birth (Malfatti et al., 2014). Interestingly, the mouse orthologs of *GUSB, NEB* and *CHRNA1* also cause pre- or perinatal lethality when knocked out in mouse, according to phenotypic data of knockout mice from Mouse Genome Informatics (MGI) (Eppig et al., 2005).

In order to systematically investigate gene essentiality in human, EGs are often inferred from the human orthologs of prenatal or preweaning lethal genes in the mouse based on the extensive similarity between the genomes of mouse and human (Dickerson et al.,

2011; Feldman et al., 2008; Georgi et al., 2013; Goh et al., 2007; Park et al., 2008). For example, using phenotypic data of knockout mice from MGI (Eppig et al., 2005), Dickerson et al. and Georgi et al. identified 1,299 and 2,472 human orthologs of lethal genes in the mouse, respectively (Dickerson et al., 2011; Georgi et al., 2013). Based on targeted mutant embryonic stem cells generated by the International Knock-out Mouse Consortium (IKMC) (Skarnes et al., 2011), the International Mouse Phenotyping Consortium (IMPC) generated and phenotyped 1,751 new knockout mouse lines on a uniform C57BL/6N background, among which 410 knockout lines displayed preweaning lethality (Dickinson et al., 2016). This study is consistent with previous observation that 30% (or ~6,000) of protein-coding genes are essential for pre- and postnatal survival in the mouse (Dickinson et al., 2016; White et al., 2013). Remarkably, the IMPC also identified 198 subviable knockout lines, which demonstrated that some genes may exhibit incomplete penetrance and variable expressivity even on a defined genetic background (Dickinson et al., 2016).

Human cell line based assays are complementary approaches to identify human EGs. RNA interference (RNAi) libraries targeting the human genome enabled earlier studies to identify cell EGs (Harborth et al., 2001; Luo et al., 2008; Silva et al., 2008). For example, Luo et al. performed RNAi screens in 12 cancer cell lines and identified 268 common EGs among the 12 cell lines (Luo et al., 2008). Recently, three genome-wide scale screens based on CRISPR/Cas9 gene editing system have been performed to assess the effect of single-gene disruption on survival of haploid human cancer cell lines (Blomen et al., 2015; Hart et al., 2015; Wang et al., 2015). These studies systematically uncovered

genes responsible for cell-based essentiality in human cell lines in a genome-wide scale. Wang et al. presented 1,878 cell essential genes in the near-haploid chronic myeloid leukemia cell line KBM7 (Wang et al., 2015). Blomen et al. identified 1,734 genes that were required for optimal growth for both KBM7 and HAP1 cell lines (Blomen et al., 2015). Hart et al. observed 1,580 genes whose perturbation decreased cell growth and proliferation in more than three studied cell lines (Hart et al., 2015). Although the core EGs discovered in these studies overlap greatly, the number of EGs identified in each study varies because different cell lines were selected and different thresholds were used for determining cell viability.

Previous studies have reported some key characteristics of EGs. Firstly, EGs tend to encode hub proteins that are most highly connected in biological networks (Goh et al., 2007; Jeong et al., 2001), showing the functional importance of EGs. Secondly, EG are often highly conserved across species. The conservation of EGs is supported by several comparative genomic studies in bacterial genomes (Bergmiller et al., 2012; Gerdes et al., 2003; Jordan et al., 2002; Luo et al., 2015) and Georgi et al., who found that EGs exhibit increased conservation across rodent and primate lineages compared to the rest of genes in the genome (Georgi et al., 2013). Thirdly, EGs are more likely to be under purifying selection, which is supported by observations that EGs were intolerant to exonic missense (Georgi et al., 2013; Petrovski et al., 2013) and loss-of-function (Lek et al., 2016) variants in general populations. Lastly, EGs are prone to exhibiting haploinsufficiency, as is suggested by both Deutschbauer et al. and Georgi et al. who found enrichment of EGs among haploinsufficient genes (Deutschbauer et al., 2005; Georgi et al., 2013).

In this chapter, I set out to compile a comprehensive list of EGs in human by combining the legacy mouse phenotyping data from MGI (Eppig et al., 2005), the newly uncovered lethal and subviable mouse genes from IMPC and human cell EGs from recent CRISPR/Cas9-based studies. With an updated set of EGs, I characterized the genetic and functional properties of EGs and confirmed the association between EGs and human disease.

Results

Identification of a comprehensive list of human orthologs of essential genes in the mouse

To identify a comprehensive list of human orthologs of EGs in the mouse, I established orthology between genes in mice and humans (Eppig et al., 2005), and used the Human Genome Mutation Database (HGMD) (Stenson et al., 2014) to annotate human disease associations. I next combined the published data from the MGI database (MP terms listed in Table 2.1) and 608 genes identified in the IMPC effort as causing lethality and subviability to compile an updated list of 3,326 EGs, along with 4,919 nonessential genes (NEGs).

The IMPC effort expanded a phenotypic spectrum for over 300 genes associated with known Mendelian diseases. From 194 subviable genes with identified human orthologs, 57 were associated with human disease, of which 34 were previously unreported for their subviable phenotypes (Supplementary Data 2.1; new reports indicated by ‘N’ in column J). For example, SET binding protein 1 (*SETBP1*) has been reported as frequently

mutated in several types of chronic leukaemia and in Schnizel-Giedion syndrome, a congenital disease characterized by a high prevalence of tumors, severe mid-face hypoplasia, heart defects and skeletal anomalies (Piazza et al., 2013; Schinzel and Giedion, 1978). Among 399 lethal genes, 126 human orthologs have been associated with human diseases, including 52 disease genes for which the IMPC effort provides the first report of their null phenotype in the mouse (Supplementary Data 2.2). The human orthologs of these novel lethal genes have been linked to metabolic and storage syndromes (*ADSL*, *DHFR*, *GYG1*, *PC*), mitochondrial complex deficiencies (*ATP5E*, *NDUFS1*, *NUBPL*, *SDHA*, *SLC25A3*, *UQCRB*), or syndromes caused by disruption of basic processes such as replication or translation initiation (*EIF2B3*, *EIF2B4*, *ORC1*). The severity of clinical manifestation of these human syndromes ranges from neonatal lethality (*BBS10*, *SLC25A3*) matching the observed phenotype in the mouse, to neurological disorders and intellectual disability (*COQ6*, *DEPDC5*, *GOSR2*, *KDM5C*, *YARS*). These differences in clinical manifestation may be due to differences between underlying biological processes in the mouse and human.

Expansion of the essential gene list through genome-wide screens for cell-essential genes in human cell lines

I used data from three recent publications on genome-wide screens for cell-essential genes in human cells to address the overlap between essential genes in the human and mouse genome (Blomen et al., 2015; Hart et al., 2015; Wang et al., 2015). I selected core essential human genes from each study and compared these to the human orthologs of mouse essential genes on the consensus list of curated IMPC-MGI genes (see Materials

and Methods). I found that approximately 35% of core essential genes in each study were associated with lethality or subviability in the mouse, and that mouse null-phenotypes for 61–62% of genes were currently unknown (Figure 2.1). Of the 19 human essential genes common to all three studies that were nonessential in the mouse, only three (*Rbmx*, *Dkc1*, and *Sod1*) could reliably be confirmed as a targeted knockout of a nonessential gene, highlighting the remarkable concordance between mouse and human in their core essential genes.

From these cell-based studies of EGs, I identified an overlapping core set of genes that were essential in the majority of cell lines tested (n = 956), but not necessarily all cell lines tested. To identify the most comprehensive set of EGs in mammals, I combined the set of human orthologs of EGs in the mouse (n =3,326) with a set of human "core EGs" (n = 956) that were found to be essential in cell-based assays (Blomen et al., 2015; Hart et al., 2015; Wang et al., 2015). Based on the significant overlap between tested mouse and human EGs, I expanded our original set of 3,326 EGs with the addition of nonoverlapping 589 EGs identified only in human cell lines for a total of 3,915 EGs (Materials and Methods, Supplementary data 2.3). In my subsequent analyses, I analyzed and compared the features of these 3,915 EGs with 4,919 human orthologs of genes with reported nonlethal phenotypes in the mouse [nonessential genes (NEGs)].

Enrichment of human disease genes and genes neighboring GWAS hits among essential genes

It has been shown that genes causing lethality in mice are enriched in human disease genes (Dickerson et al., 2011; Georgi et al., 2013). With these updated EG lists, I report

an even stronger enrichment of essential genes relative to nonessential for human disease genes catalogued in the HGMD (odds ratio = 2.00, P = 6.83×10^{-39}, Figure 2.2A). Consistent with this enrichment, of the 3,302 protein-coding HGMD disease genes, 2,434 have a reported phenotype and more than half (1,253) are essential in mice (Figure 2.2B; Supplementary Data 2.3). Furthermore, I found an enrichment of EGs in comparison to nonessential genes (odds ratio = 1.16, P value = 0.0015) among 6,384 genes encompassing or neighboring the disease- and trait-associated variants in the NHGRI-EBI catalogue of published genome-wide association studies ('GWAS hits') (Welter et al., 2014) (Figure 2.3).

Essential genes' intolerance to deleterious mutations

With the updated catalog of mouse 3,326 EGs and 4,919 NEGs, I compared the mutability of their human orthologs in exome sequences of 60,706 subjects in the Exome Aggregation Consortium data (ExAC, http://exac.broadinstitute.org) (Lek et al., 2016). The ExAC data were used to generate intolerance scores for all protein-coding genes by two complementary methods: a) the residual variation intolerance score (RVIS), which is based on intolerance to common missense and truncating single nucleotide variation (http://genic-intolerance.org/) (Petrovski et al., 2013); and b) the estimation of probability of being loss-of-function intolerant (pLI score) (Lek et al., 2016). Human orthologs of EGs are more intolerant to variation (low RVIS and high pLI scores) than orthologs of NEGs and all genes in the human genome (P value < 2.2×10^{-16} for lower percentiles in essential genes using the two scoring systems, Figure 2.4A, B). Moreover, the IMPC effort identified a set of 22 human orthologs of EGs that were not previously associated

with human disease (Figure 2.5, Table 2.2), but based on their intolerance to functional variation and lethality of their null alleles in the mouse, they represent strong candidates for undiagnosed human diseases.

Chromosomal distribution of essential genes

EGs are distributed throughout the human genome, with the exception of the Y chromosome (Figure 2.6). I identified 3 chromosomal regions, 17q21, 11q13 and 16q22, with significant enrichment of EGs (FDR<0.05; one-sided Fisher's exact test) (Supplementary data 2.4). 17q21 contains 64 EGs which collectively cover 14.0% (1.69 Mb) of the total length of the region. 11q13 contains 63 EGs which collectively cover 15.5% (2.13 Mb) of the total length of the region. 16q22 contains 37 EGs which collectively cover 29.3% (2.17 Mb) of the total length of the region. I further annotated these three chromosomal regions with associated human diseases from OMIM (Hamosh et al., 2005) and identified 91 human disease loci (including frontotemporal dementia, progressive myoclonic epilepsy-6 and mental retardation) associated with 17q21, 59 disease loci associated with 11q13, as well as 30 disease loci associated with 16q22 (Supplementary data 2.4). Interestingly, 17q21 has been shown by replicated genome-wide linkage studies (Cantor et al., 2005; Yonan et al., 2003) to harbor susceptibility to autism spectrum disorder (ASD). In the mouse genome, I identified 2qB as the only chromosomal region with significant enrichment of EGs in the mouse (44 EGs vs. 21 NEGs; P value =0.048; One-sided Fisher's exact test after Bonferroni correction) (Figure 2.7, Supplementary data 2.5).

Disease categories associated with essential genes

To systematically evaluate the relevance of EGs in diseases categorized by affected tissues and age of onset, I obtained lists of human disease genes annotated by the Human Phenotype Ontology (Kohler et al., 2017). First, I tested for enrichment of 3,915 EGs vs. 4,919 NEGs among 24 genes sets associated with abnormality of different organs or systems. Except for one disease category ("Abnormality of the thoracic cavity"), all of the gene sets associated with the rest of the 23 disease categories were significantly enriched for EGs (Table 2.3), which suggests that disturbance of EGs may contribute to a wide variety of diseases affecting different organs or systems. When genes were categorized by the age of onset of their associated diseases, I observed that EGs are significantly enriched among disease genes annotated as "congenital onset" (at birth) (Odds ratio=3.84, P value=6.59×10^{-13}; Two sided Fisher's exact test), "neonatal onset" (within 28 days) (Odds ratio=3.15, P value=0.015) and "infantile onset" (between 28 days and 1 year) (Odds ratio=2.44, P value=8.29×10^{-11}), but not among genes annotated as "childhood onset" (between 1 year and 5 years) (Odds ratio=1.01, P value=1), "juvenile onset" (between 5 years and 15 years) (Odds ratio=1.16, P value=0.56) or "adult onset" (16 years or later) (Odds ratio=1.43, P value=0.083) (Figure 2.8). These results suggest that EGs may play a distinct role in early on-set diseases.

Expression patterns of essential genes across tissues

To evaluate the tissue specificity and ubiquitousness of EG expression, we analyzed the expression patterns of EGs and NEGs over multiple human tissues using transcriptomic data from GTEx (The GTEx Consortium, 2015). Compared to NEGs, a higher proportion

of EGs are ubiquitously expressed and a lower proportion of EGs are specifically expressed in certain tissues (Figure 2.9). Among 6,455 ubiquitously expressed genes (entropy score>=5.5), there were 1,477 EGs and 941 NEGs, representing a significant enrichment of EGs (Two-sided Fisher's exact test: p-value=2.10×10^{-134}, Odds ratio=3.42) (Supplementary data 2.3). Among 1,680 genes specifically expressed in certain tissues (entropy score<=1.0), there were 116 EGs and 415 NEGs, representing a significant depletion of EGs (Two-sided Fisher's exact test: p-value=1.58×10^{-29}, Odds ratio=0.33). The top 5 tissue types containing tissue-specifically expressed EGs were testis (with 27 EGs), liver (with 15 EGs), muscle (with 12 EGs), kidney (with 8 EGs) and brain (with 7 EGs) (Table 2.4).

Haploinsufficiency of essential genes

Homozygous loss-of-function mutations in EGs lead to lethality (or miscarriages in humans) and as such, cannot contribute to disease. Although a depletion of loss-of-function mutations in EGs in humans was reported (Georgi et al., 2013; Petrovski et al., 2013), heterozygosity for a loss-of-function mutation or other "milder" alleles in EGs may contribute to both dominant and recessive diseases. I illustrate this point using a catalog of disease-linked genes in Online Mendelian Inheritance in Man (Hamosh et al., 2005). EGs were enriched relative to NEGs in 1,000 genes underlying dominant diseases (odds ratio = 1.95, P value = 3.17×10^{-19}; two-sided Fisher's exact test) and 1,645 genes underlying recessive disease (odds ratio = 1.52, P value = 4.94×10^{-11}; two-sided Fisher's exact test) (Figure 2.10). A stronger enrichment of EGs among genes underlying dominant disease compared to recessive disease implies that dominant negative alleles

and haploinsufficiency play an important role. I provide multiple lines of evidence for higher probability of haploinsufficiency of EGs (Figure 2.10). First, using the systematically rated dosage-sensitive genes from ClinGen (Rehm et al., 2015), I found that EGs were significantly enriched compared with NEGs and that the levels of EG enrichment positively correlated with levels of evidence supporting dosage sensitivity of rated genes (odds ratio = 3.94, P value = 5.07×10^{-20} for "sufficient evidence"; odds ratio = 5.26, P value = 7.08×10^{-5} for "some evidence"; odds ratio = 2.52, P value = 0.0106 for "little evidence"; odds ratio = 1.14, P value = 0.608 for "not dosage sensitive"; two-sided Fisher's exact test). Second, as an extension of the earlier findings from the work by Georgi et al. (Georgi et al., 2013), I confirmed the enrichment of EG relative to NEG for 262 human haploinsufficient genes (Dang et al., 2008) with the updated EG and NEG list (183 EGs vs. 62 NEGs; P value = 1.64×10^{-22}, odds ratio = 3.84; two-sided Fisher's exact test). Third, EGs are significantly overrepresented among 313 human orthologs of mouse genes with heterozygous alleles associated with mutant phenotypes from the MGI (Eppig et al., 2005) (odds ratio = 3.43, P value = 2.74×10^{-23}; two-sided Fisher's exact test). Fourth, with two genome-wide prediction models of haploinsufficient genes in the human genome (Huang et al., 2010; Steinberg et al., 2015), I observed that EGs have significantly higher probability of exhibiting haploinsufficiency compared with NEGs (P value < 2.2×10^{-16} for both models; two-sided Wilcoxon rank sum test) (Figure 2.11 A and B). Based on the findings that EGs linked to Mendelian disease are overwhelmingly dosage-sensitive, in Chapter 3 I explored the possibility that a cumulative effect of pathogenic variants in multiple EGs may underlie the genetic basis of a complex disease with early postnatal onset, such as ASD.

Discussion

I complied the most comprehensive EG set established to date (n=3,915) by combining phenotypic data in knockout mice (n=3,326) and data from genomic-scale human cell assays (n=956). I confirmed the important role of EGs in human disease by showing that EGs comprise a major part of disease genes and that EGs are enriched among genes neighboring GWAS hits. While EGs are distributed throughout the genome (with the exception of the Y chromosome) and tend to be ubiquitously expressed across different tissues, I identified three EG-enriched chromosomal regions, among which 17q21 was associated with ASD according to replicated genome-wide linkage studies (Cantor et al., 2005; Yonan et al., 2003). Finally, with an updated EG set, I confirmed that EGs are intolerant to deleterious mutations and are more likely to be haploinsufficient.

The current catalog of human EGs includes 3,915 genes, which is a substantial increase since the publication by Georgi et al. in 2013 (n=2,472) (Georgi et al., 2013). Based on studies in the mouse, 30% of genes in a mammalian genome are essential (Dickinson et al., 2016; White et al., 2013), meaning that the current catalog includes more than 65% of the core set of “the indispensable genome”. With a major portion of EGs identified, the general functional and genetic properties of EGs can be credibly characterized, as is shown in this chapter. There is still a great amount work to be done to identify the complete set of EGs. Systematically generating and phenotyping knockout mice for every gene in the mouse genome (~20,000 genes) is a feasible strategy to achieve this goal, which has been one of the main objectives of the International Mouse Phenotyping Consortium (IMPC) (Dickinson et al., 2016; Koscielny et al., 2014).

Recent breakthroughs in human cell line based assays on cell proliferation and survival provided an effective alternative way to identify EGs in human (Blomen et al., 2015; Hart et al., 2015; Wang et al., 2015). It overcomes some limitations of the knockout mice based approach such as biological and genomic differences between mouse and human. Indeed, some genes that are essential in one organism may not be essential in other organisms, as is suggested by a few comparative genomic studies in bacterial genomes (Bergmiller et al., 2012; Gerdes et al., 2003). However, the cell line based approaches have their own drawbacks. Firstly, the precise number of cell EGs is difficult to determine, since it depends on the chosen threshold for impaired fitness of cell lines (Wang et al., 2015). Secondly, most of the genomic screens have been performed on human cancer cell lines with gene knockouts. Cases of discrepancy may occur when inferring organismal lethality from cancer cell survival rate. For example, the cancer cell line based assays can reveal oncogenes (Luo et al., 2008) which may not necessarily be essential in non-cancer cell lines. Therefore, careful examination and comparison of EGs inferred from lethality of knockout mice and viability of human cell lines is warranted as the catalog of known EGs grows continuously.

In contrast to earlier studies which suggested that human disease genes tend to be non-essential (Domazet-Loso and Tautz, 2008; Feldman et al., 2008; Goh et al., 2007; Park et al., 2008), my results are consistent with the conclusions from Dickerson et al. (Dickerson et al., 2011) and Georgi et al. (Georgi et al., 2013) that a major portion of human disease genes are essential. The enrichment of EGs among human disease genes from HGMD (Stenson et al., 2014) and genes neighboring GWAS hits (Welter et al.,

2014) suggests that mutations in EGs contribute not only to Mendelian diseases, but also to complex traits and disorders. My results on the enrichment of EGs among different types of disease genes showed that while the contribution of EGs is widespread across disorders with various affected systems and underlying mechanisms, EGs seems to play an especially important role in early onset diseases. Homozygous loss-of-function mutations in EGs are not present in living individuals. While EGs are generally intolerant to other deleterious alleles at the population level, because of the functional importance of EGs, observed deleterious alleles in EGs are more likely to be pathogenic. I confirmed that EGs tend to demonstrate haploinsufficiency and an autosomal dominant model of inheritance (Dickerson et al., 2011; Georgi et al., 2013), which supports a potential cumulative effect of deleterious mutations (mostly in heterozygous state) in EGs on the risk of early onset complex disorders, such as ASD.

Materials and Methods

Identification of essential genes and non-essential genes

I identified 3,023 protein-coding EGs annotated with 50 mouse phenotype (MP) terms, including prenatal, perinatal, and postnatal lethal phenotypes from the MGI (Eppig et al., 2005) (Table S8). The MGI database was also used to extract 4,995 protein-coding NEGs with nonlethal phenotypes in the mouse. Phenotype data from the IMPC database portal (Koscielny et al., 2014) expanded the lethal gene list with the addition of 252 lethal genes and 101 genes with subviable phenotypes. I further supplemented the nonlethal gene list with 701 genes with viable phenotypes from the IMPC. In the case of discrepancy in the

reported lethality status between the MGI and the IMPC, I deferred to the phenotypes reported by the IMPC, because these mouse lines were generated on a defined C57BL/6N background and phenotypically characterized using a standardized pipeline.

One to one mouse–human orthology of lethal and nonlethal genes was established based on MGI annotation and manual curation, resulting in 3,326 essential and 4,919 nonessential human orthologs (NEGs). The catalog of EGs was further augmented with the addition of cell EGs from three recent studies (Blomen et al., 2015; Hart et al., 2015; Wang et al., 2015) aimed at the characterization of EGs in human cell lines. I obtained 1,580 core EGs (genes above essentiality threshold in at least three of five cell lines in the study) from the work by Hart et al. (Hart et al., 2015), 1,739 core EGs (genes above essentiality threshold in at least two of four cell lines in the study) from the work by Wang et al. (Wang et al., 2015), and 1,734 core EGs (genes above essentiality threshold in at least one of two cell lines in the study) from the work by Blomen et al. (Blomen et al., 2015). By taking the overlap of three sets of core EGs, I obtained 956 high-confidence human EGs. Among 956 EGs in human cell lines, 348 genes (36.4%) are also human orthologs of EGs in the mouse, 19 genes (2.0%) are human orthologs of NEGs in the mouse, and 589 genes (61.6%) have not been tested in the mouse.

Identification of genes encompassing or surrounding disease- and trait-associated SNPs (‘GWAS hits’)

6,384 protein-coding genes encompassing and/or neighboring disease- or trait-associated variants (‘GWAS genes’) were obtained from the GWAS Catalog (Welter et al., 2014)

(downloaded on April 29, 2016). Specifically, I used the 'mapped genes' from the GWAS Catalog, which are defined as the genes mapped to the strongest SNP from GWAS reports. The mapped genes are defined as the genes encompassing the GWAS SNP(s), (that is, located in coding or intragenic regions; n = 4,228) or the two genes that map upstream and downstream of the GWAS SNP(s) (that is, in intergenic regions; n = 3,422). Enrichment of GWAS genes between our gene sets of interest was assessed by two-sided Fisher's exact test.

Categorization of human disease genes by the Human Phenotype Ontology

I categorized human disease genes collected by the Human Phenotype Ontology Project (accessed the February 2017 release) (Kohler et al., 2017) based the subclasses of two ontology terms: "Phenotypic abnormality" (HP:0000118) and "Onset" (HP:0003674). As a result, I obtained 24 sets of disease genes annotated with abnormality in different organs or systems (i.e. "Abnormality of prenatal development or birth"," Abnormality of the breast"," Abnormality of the musculature"," Abnormality of limbs"," Abnormality of the voice"," Growth abnormality"," Abnormality of the respiratory system"," Abnormality of the skeletal system"," Abnormality of head or neck"," Abnormality of the digestive system"," Abnormality of the cardiovascular system"," Abnormality of the eye"," Abnormality of connective tissue"," Abnormality of the genitourinary system"," Neoplasm"," Abnormality of the nervous system"," Abnormality of the ear"," Abnormality of the integument"," Abnormality of the immune system"," Abnormality of blood and blood-forming tissues"," Abnormality of the endocrine system"," Abnormality of metabolism/homeostasis"," Constitutional symptom" and" Abnormality of the thoracic

cavity") , and 6 sets of genes annotated with different age of onset for their associated diseases (i.e. "congenital onset", "neonatal onset", "infantile onset", "childhood onset", "juvenile onset" and "adult onset"). For each set of disease genes, I evaluated the enrichment of 3,915 EGs vs. 4,919 NEGs using two sided Fisher's exact test. For a 2 * 2 contingency table with 4 cell counts: *a* (# EGs in target gene set), *b* (# EGs not in target gene set), *c* (# NEGs in target gene set) and *d* (# NEGs not in target gene set), the 95% confidence interval of odds ratio (OR) is calculated as follows:

$$exp[ln(OR) \pm 1.96 * \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]$$

Chromosomal distribution of essential genes

The chromosomal distribution of 3,882 EGs with available coordinates in genome build hg19 was plotted using Phenogram (http://visualization.ritchielab.psu.edu/phenograms/plot). The enrichment of EGs across cytobands was assessed by Enrichr (Chen et al., 2013).

Tissue specificity and ubiquitousness of gene expression

For each expression dataset, we measure the Shannon entropy score as suggested by Schug et al.(Schug et al., 2005). The definition of tissue specificity score is shown below:

Suppose the expression levels of relevant genes were measured in *N* tissues in an expression dataset, the relative expression level of gene *g* in tissue *t* was defined as:

$$p_{t|g} = w_{g,t} / \sum_{1 \le t \le N} w_{g,t}$$

where $w_{g,t}$ is the measured expression level of gene *g* in tissue *t*.

The Shannon entropy (*H*) measuring the distribution of expression levels of gene *g* across all tissues in an expression dataset was defined as:

$$H_g = \sum_{1 \le t \le N} - p_{t|g} \log_2(p_{t|g})$$

The tissue specificity score of gene *g* in tissue *t* is calculated as:

$$Q_{g|t} = H_g - \log_2(p_{t|g})$$

Analysis of haploinsufficiency of essential genes

I collected genes sets from multiple studies and resources for the analysis of patterns of inheritance and haploinsufficiency of EGs. First, a catalog of human disease genes was obtained from Online Mendelian Inheritance in Man (OMIM; downloaded on July 12, 2016) (Hamosh et al., 2005). From the OMIM catalog, I identified 1,411 genes annotated with genetic disorders that are “autosomal dominant” or “X-linked dominant” and 2,056 genes annotated with genetic disorders that are “autosomal recessive” or “X-linked recessive.” By dissecting the above two gene lists, I obtained 1,000 genes underlying only dominant diseases, 1,645 genes underlying only recessive diseases, and 441 genes that were linked to both dominant and recessive disorders. Second, a list of 616 protein-coding genes that were systematically assessed for evidence for dosage sensitivity was obtained from ClinGen Dosage Sensitivity Map (Rehm et al., 2015). Among 616 genes, 239 genes were dosage-sensitive with sufficient evidence, 41 genes were dosage-sensitive with some evidence, 47 genes were dosage-sensitive with little evidence, 200

genes had no evidence for dosage pathogenicity so far, and 89 genes were not dosage-sensitive or with autosomal recessive phenotype. Third, a list of 262 haploinsufficient genes based on textmining from PubMed and OMIM was obtained from the work by Dang et al. (Dang et al., 2008). Fourth, from the MGI, I identified 313 human orthologs of mouse genes associated with heterozygous phenotypes. For each of the gene sets, I evaluated the enrichment of EGs compared with NEGs using Fisher's exact test.

I acquired the Haploinsufficiency Scores (Huang et al., 2010) and the Genome-Wide Haploinsufficiency Scores (Steinberg et al., 2015) for genome-wide prediction of the probability of haploinsufficiency. For each prediction model, the raw scores were ranked and converted to percentiles. The histograms and estimated density curves were plotted using ggplot2 (geom_histogram and geom_line) in R.

Figures

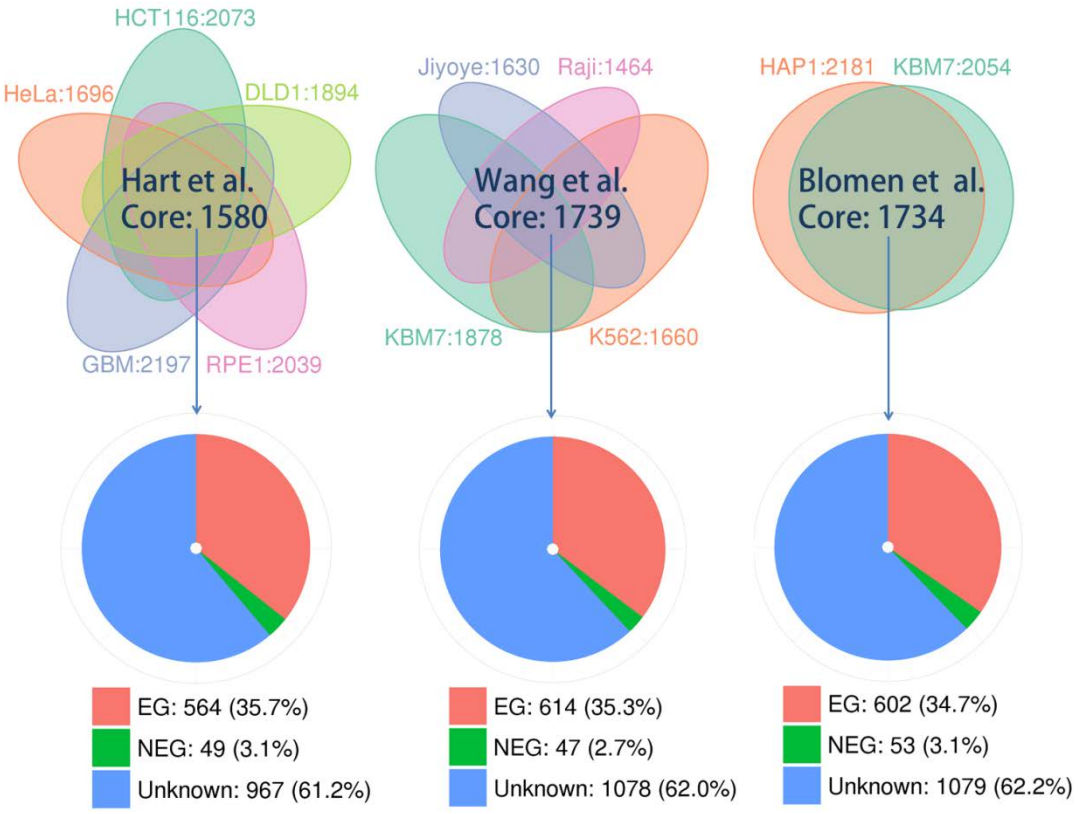

Figure 2.1 Overlap between essential genes in human cells and human orthologs of essential genes in the mouse. Core essential genes in human cells identified in three studies: 1,580 (Hart et al., 2015), 1,739 (Wang et al., 2015), and 1,734 (Blomen et al., 2015) (top row) (see Methods). Pie charts indicate overlap between core human cell-essential genes and orthologous genes in the mouse: essential (EG, red); nonessential (NEG, green) and genes with unknown function in the mouse (Unknown, blue).

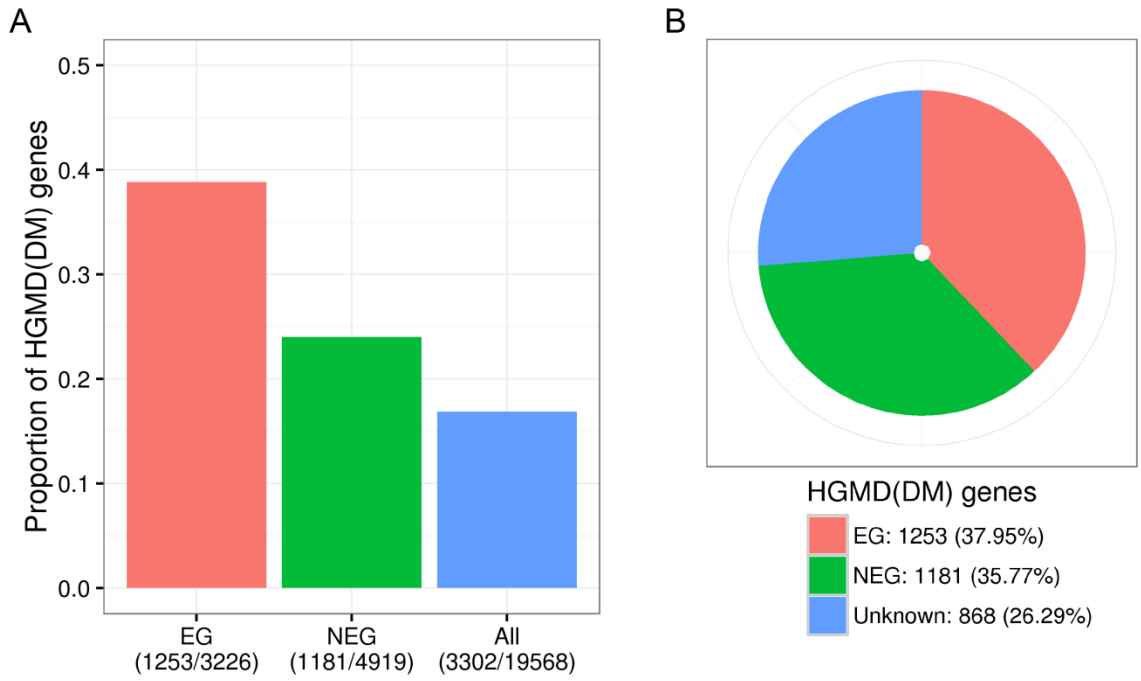

Figure 2.2 Enrichment of essential genes among HGMD human disease genes. (A) The fractions indicate the number of HGMD disease genes (disease-causing mutations (DM)) (n = 3,302) among 3,326 essential genes (EG, red); 4,919 nonessential genes (NEG, green) and 19,568 protein-coding genes (All, blue). Fisher's exact test for enrichment: EG versus NEG (odds ratio = 2.00, $P = 7.80 \times 10^{-46}$), EG versus All (odds ratio = 3.13, $P = 2.42 \times 10^{-160}$), NEG versus All (odds ratio = 1.56, $P = 1.83 \times 10^{-29}$). **(B)** Essentiality status of 3,302 HGMD disease genes.

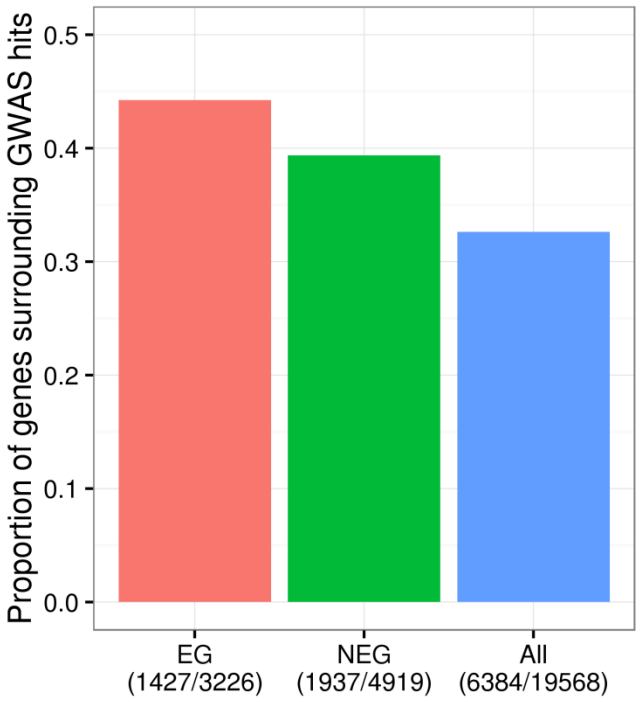

Figure 2.3 Enrichment of essential genes among genes neighboring GWAS hits. The fractions indicate the number of genes encompassing or neighboring GWAS hits (Welter et al., 2014) (n = 6,384) divided by essentiality status (EG in red, NEG in green, All in blue). Fisher's exact test for enrichment: EG versus NEG (odds ratio = 1.16, P = 0.0015), EG versus All (odds ratio = 1.56, $P = 5.80 \times 10^{-31}$), NEG versus All (odds ratio = 1.35, $P = 1.18 \times 10^{-19}$).

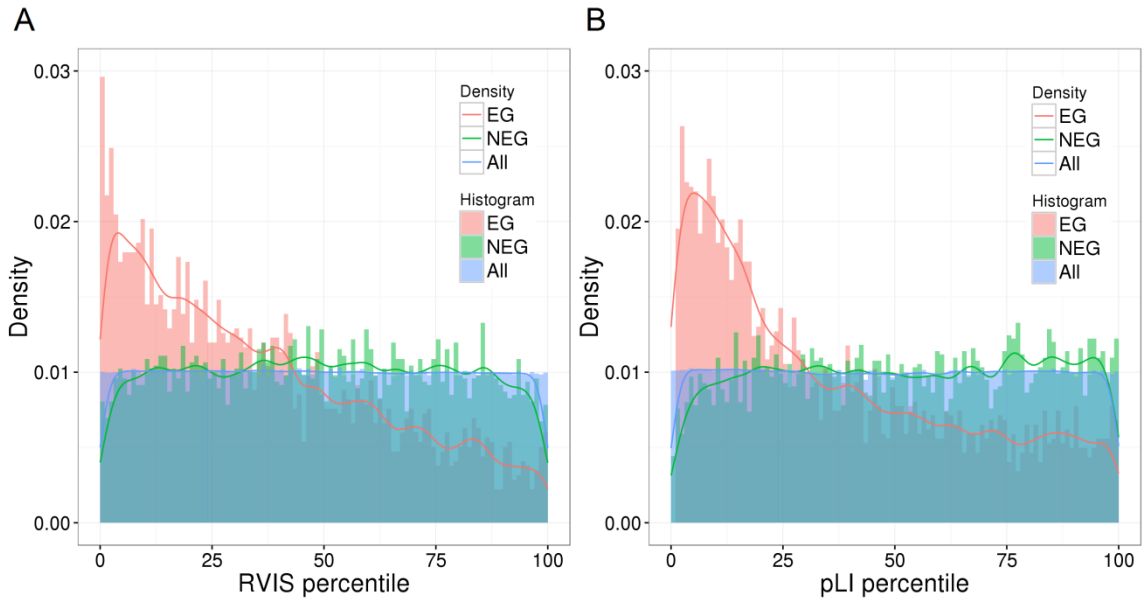

Figure 2.4 Essential genes are intolerant to deleterious mutations. (A) Distribution of percentiles of the residual variation intolerance score (RVIS) across three classes of genes: EG (in red), NEG (in green) and All (in blue). Wilcoxon rank-sum test: EG versus NEG (P value < 2.2×10^{-16}), EG versus All (P value < 2.2×10^{-16}), NEG versus All (P value = 0.579). **(B)** Distribution of percentiles of the probability of being loss of function intolerant (pLI) across three classes of genes: EG (in red), NEG (in green) and All (in blue). Wilcoxon rank-sum test: EG versus NEG (P value < 2.2×10^{-16}), EG versus All (P value < 2.2×10^{-16}), All versus NEG (P value = 4.15×10^{-5}).

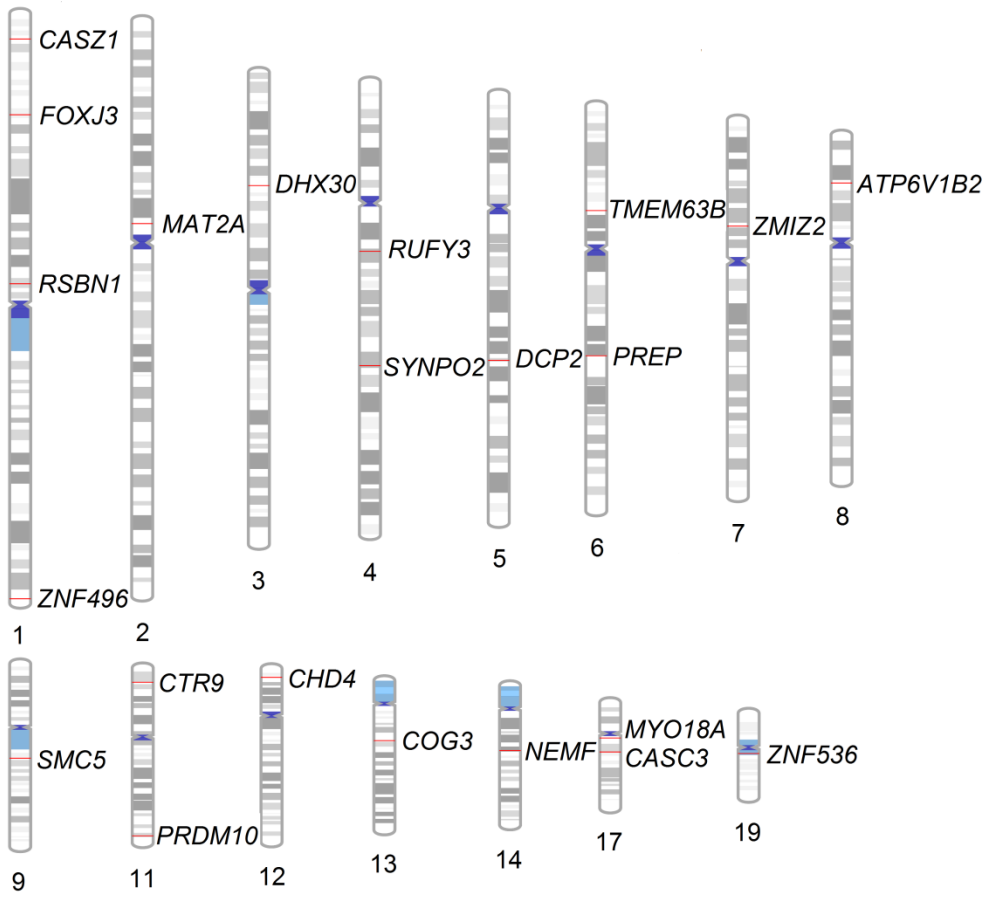

Figure 2.5 Chromosomal distribution of 22 human orthologs of mutation-intolerant essential genes that are not currently included in the catalogs of Mendelian disease genes. Red bars indicate the chromosomal positions of the exhibited genes.

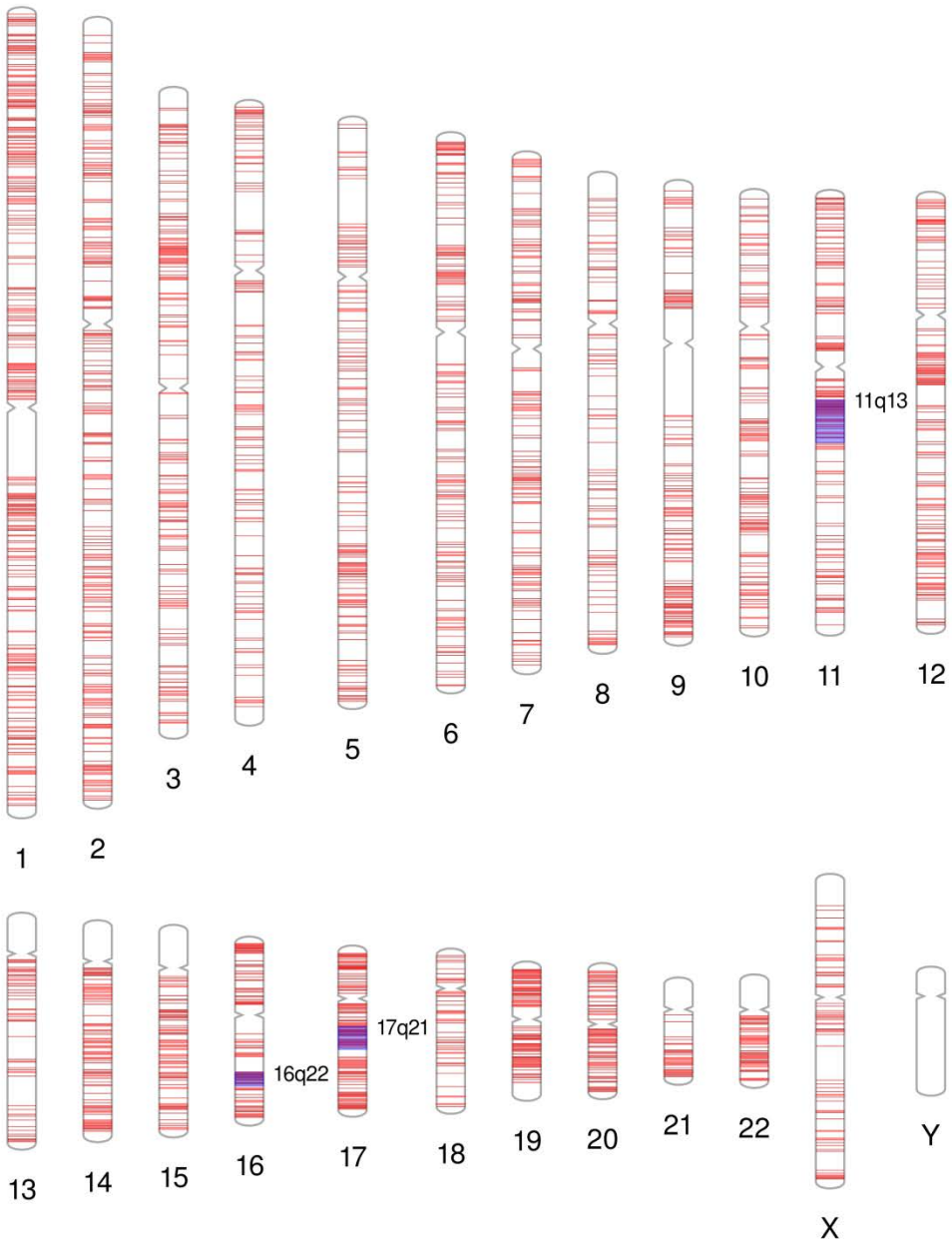

Figure 2.6 Chromosomal distribution of 3,915 human essential genes. Chromosomal positions of EGs (hg19) are shown in red. Three chromosomal regions (17q21, 11q13 and 16q22) with significant enrichment of EGs are shown in blue.

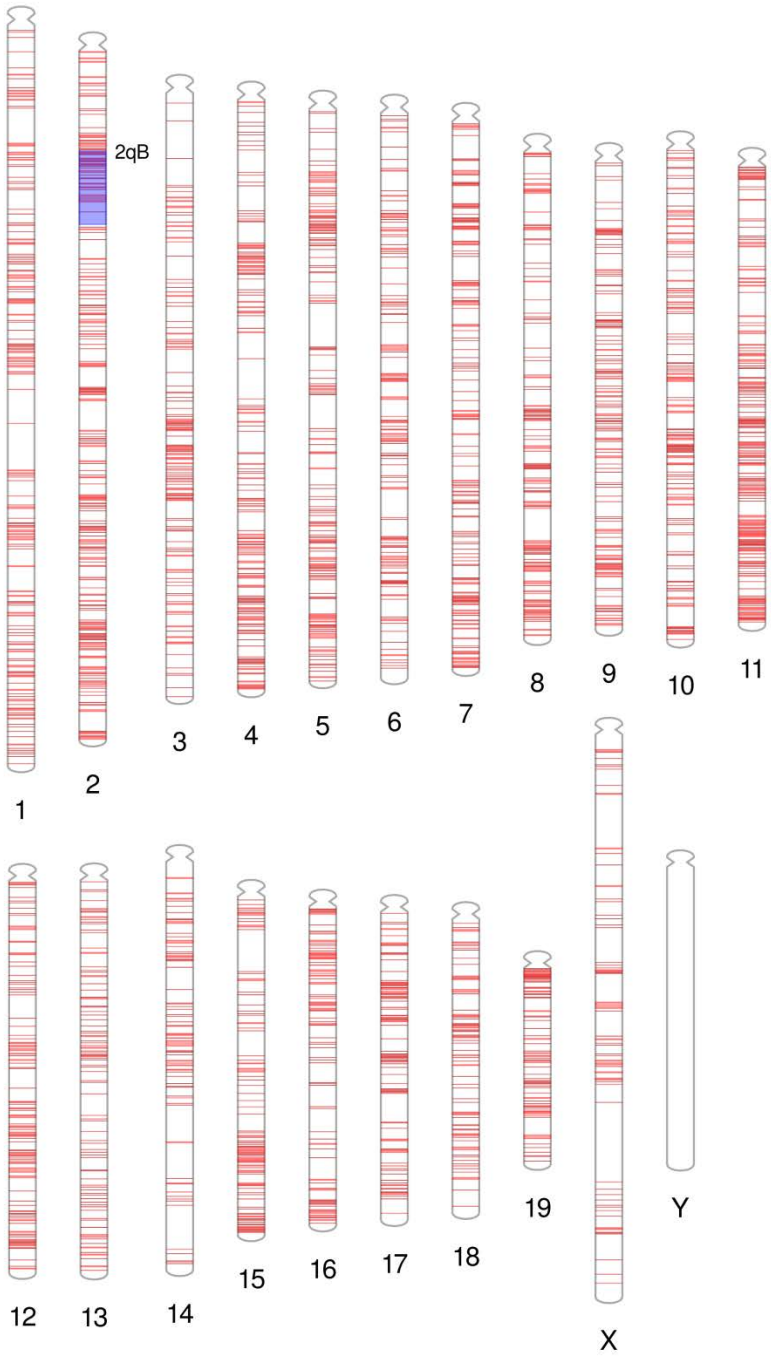

Figure 2.7 Chromosomal distribution of 3,879 essential genes in the mouse.

Chromosomal positions of EGs (mm10) are shown in red. The chromosomal region (2qB) with significant enrichment of EGs is shown in blue.

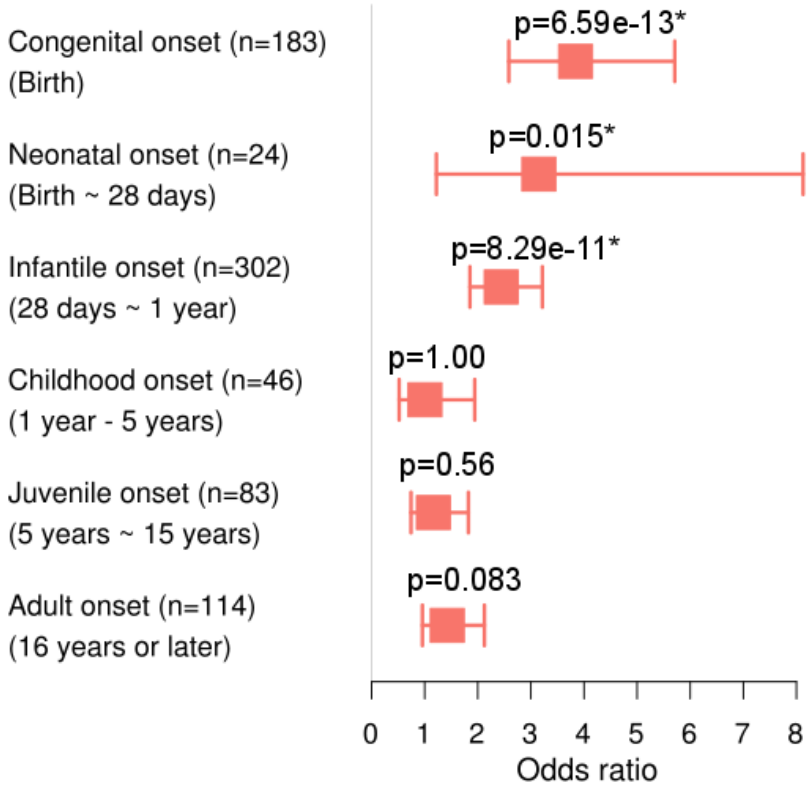

Figure 2.8 Essentiality statuses of human diseases genes categorized by age of onset. The results (p values, odds ratios and 95% confidence intervals of odds ratios) of Fisher's exact tests of enrichment of 3,915 EGs vs. 4,919 NEGs for human disease genes categorized by the age of onset of associated diseases were plotted.

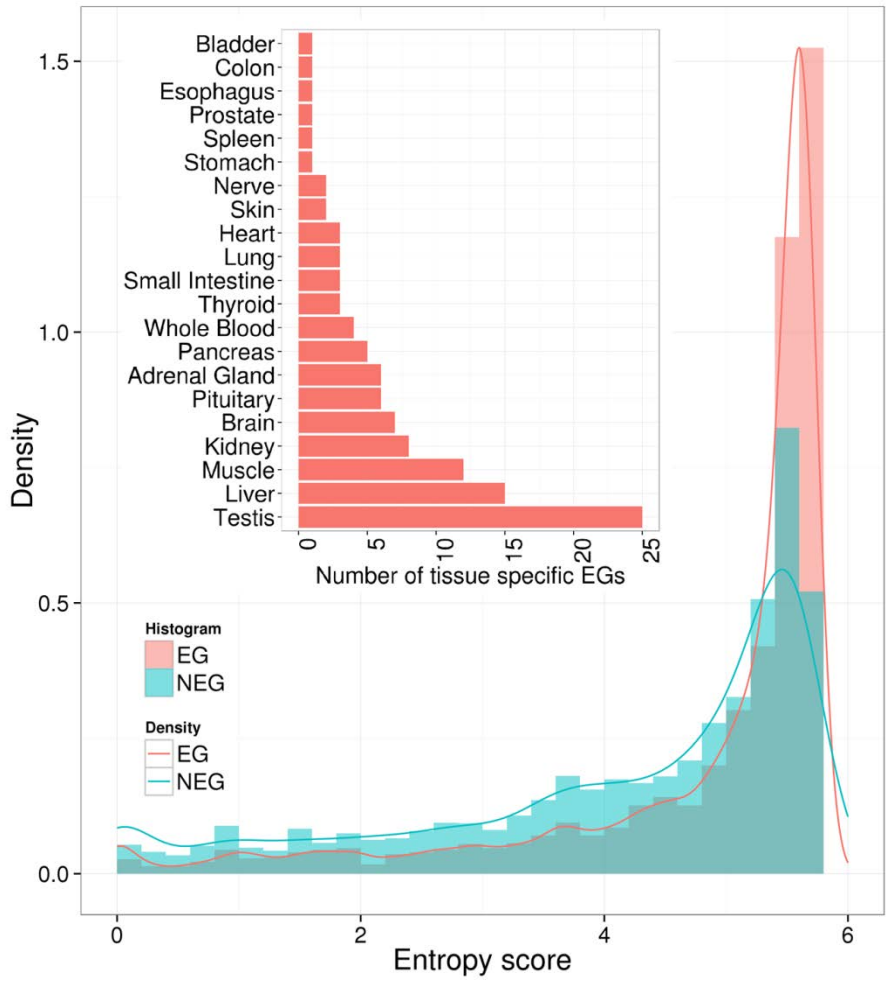

Figure 2.9 Tissue expression specificity of EGs and NEGs. The distribution of Shannon entropy scores (Methods and Materials) for EGs (in red) and NEGs (in turquoise) was shown as both histogram and estimated density curve. The numbers of tissue-specific (entropy score<1.0) EGs are shown in the top bar plot.

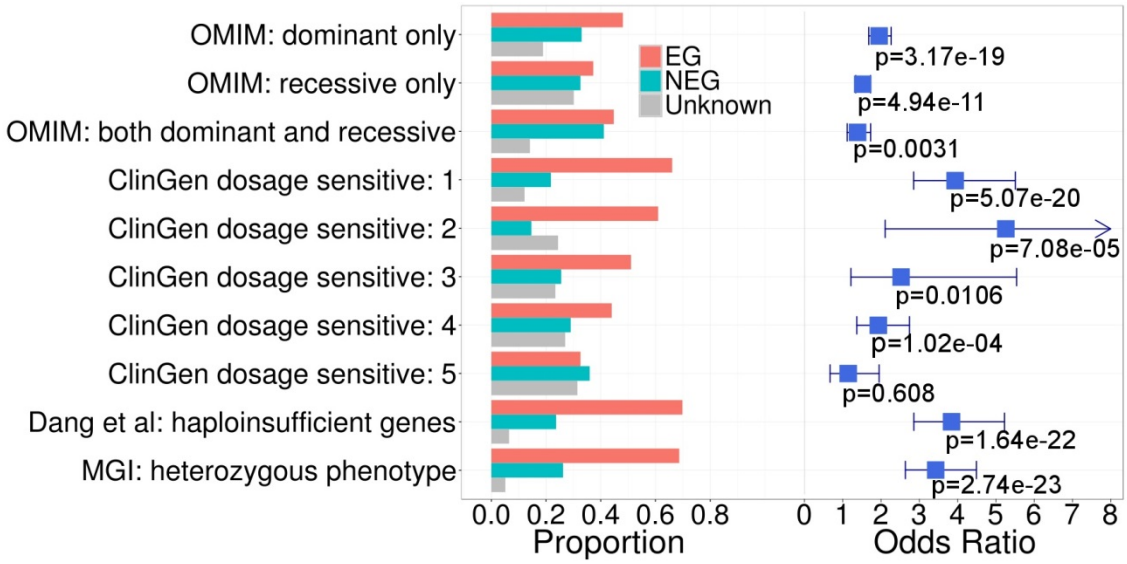

Figure 2.10 Haploinsufficiency of essential genes. For each class of genes with different essentiality status (EG in red, NEG in turquoise, and unknown in gray), the proportion of genes among each gene set of interest is plotted in *Left*. Dosage-sensitive genes from ClinGen (Rehm et al., 2015) were classified into five categories (1, sufficient evidence; 2, some evidence; 3, little evidence; 4, no evidence and 5, not sensitive/recessive). Two-sided Fisher's exact test was performed to assess the enrichment of EGs vs. NEGs, and the P values were indicated. The odds ratios for enrichment of EGs compared with NEGs and the 95% confidence intervals of odds ratios are plotted in Right. OMIM, Online Mendelian Inheritance in Man (Hamosh et al., 2005).

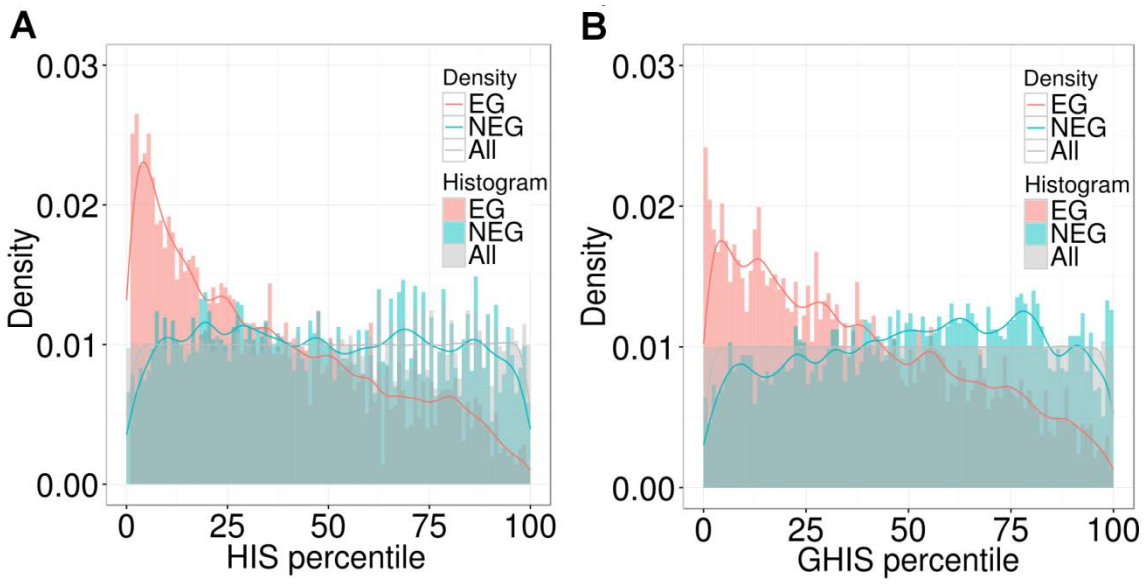

Figure 2.11 Distribution of genome-wide haploinsufficiency scores. Histograms and estimated density curves indicating the distribution of **(A)** the Haploinsufficiency Score (HIS) (Huang et al., 2010) and **(B)** the Genome-Wide Haploinsufficiency Score (GHIS) (Steinberg et al., 2015) across three gene sets, including EGs (red), NEGs (turquoise), and all protein-coding genes (56) (gray). EGs have significantly higher probability of exhibiting haploinsufficiency compared with NEGs (P value < 2.2×10^{-16} for both models; two-sided Wilcoxon rank sum test).

Tables

Table 2.1 Mouse phenotype (MP) terms for lethal phenotypes.

MP ID	Lethality type	Lethality description
MP:0002058	neonatal lethality	death within the neonatal period after birth (Mus: P0)
MP:0002080	prenatal lethality	death anytime between fertilization and birth (Mus: approximately E18.5)
MP:0002081	perinatal lethality	death anytime within the perinatal period (Mus: E18.5 through postnatal day 1)
MP:0002082	postnatal lethality	premature death anytime between the neonatal period and weaning age (Mus: P1 to approximately 3 weeks of age)
MP:0006204	embryonic lethality before implantation	death anytime between fertilization and implantation (Mus: E0 to less than E4.5)
MP:0006205	embryonic lethality between implantation and somite formation	death anytime between the point of implantation and somite formation (Mus: E4.5 to less than E8)
MP:0006206	embryonic lethality between somite formation and embryo turning	death anytime between somite formation and the initiation of embryo turning (Mus: E8 to less than E9)
MP:0006207	embryonic lethality during organogenesis	death anytime between embryo turning and the completion of organogenesis (Mus: E9-9.5 to less than E14)
MP:0006208	lethality throughout fetal growth and development	death anytime between the completion of organogenesis and birth (Mus: E14 to approximately E18.5)
MP:0008527	embryonic lethality at implantation	death due to failure of implantation (Mus: E4.5)
MP:0008569	lethality at weaning	premature death at weaning age, often due to the inability to make the transition to solid food
MP:0008762	embryonic lethality	death of an animal within the embryonic period prior to organogenesis (Mus: prior to E14)
MP:0009850	embryonic lethality between implantation and placentation	death anytime between the point of implantation and the initiation of placentation (Mus: E4.5 to less than E9)
MP:0010770	preweaning lethality	death anytime between fertilization and weaning age (Mus: approximately 3-4 weeks of age)
MP:0010831	partial lethality	the appearance of lower than Mendelian ratios of offspring of a given genotype due to death of some, but

		not all of the organisms
MP:0010832	lethality during fetal growth through weaning	death anytime between the completion of organogenesis and weaning age (Mus: E14 to approximately 3 weeks of age)
MP:0011083	complete lethality at weaning	premature death at weaning age of all organisms of a given genotype in a population, often due to the inability to make the transition to solid food
MP:0011084	partial lethality at weaning	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms at weaning age
MP:0011085	complete postnatal lethality	premature death anytime between the neonatal period and weaning age of all organisms of a given genotype in a population (Mus: P1 to approximately 3 weeks of age)
MP:0011086	partial postnatal lethality	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms anytime between the neonatal period and weaning age (Mus: P1 to approximately 3 weeks of age)
MP:0011087	complete neonatal lethality	death of all organisms of a given genotype in a population within the neonatal period after birth (Mus: P0)
MP:0011088	partial neonatal lethality	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms within the neonatal period after birth (Mus: P0)
MP:0011089	complete perinatal lethality	death of all organisms of a given genotype in a population within the perinatal period (Mus: E18.5 through postnatal day 1)
MP:0011090	partial perinatal lethality	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms within the perinatal period (Mus: E18.5 through postnatal day 1)
MP:0011091	complete prenatal lethality	death of all organisms of a given genotype in a population between fertilization and birth (Mus: approximately E18.5)
MP:0011092	complete embryonic lethality	death of all organisms of a given genotype in a population within the embryonic period prior to organogenesis (Mus: prior to E14)
MP:0011093	complete embryonic lethality at implantation	death of all organisms of a given genotype in a population at the point of implantation (Mus: E4.5)
MP:0011094	complete embryonic lethality before	death of all organisms of a given genotype in a population between fertilization and implantation (Mus: E0 to less

	implantation	than E4.5)
MP:0011095	complete embryonic lethality between implantation and placentation	death of all organisms of a given genotype in a population between the point of implantation and the initiation of placentation (Mus: E4.5 to less than E9)
MP:0011096	complete embryonic lethality between implantation and somite formation	death of all organisms of a given genotype in a population between the point of implantation and somite formation (Mus: E4.5 to less than E8)
MP:0011097	complete embryonic lethality between somite formation and embryo turning	death of all organisms of a given genotype in a population between somite formation and the initiation of embryo turning (Mus: E8 to less than E9)
MP:0011098	complete embryonic lethality during organogenesis	death of all organisms of a given genotype in a population between embryo turning and the completion of organogenesis (Mus: E9-9.5 to less than E14)
MP:0011099	complete lethality throughout fetal growth and development	death of all organisms of a given genotype in a population between the completion of organogenesis and birth (Mus: E14 to approximately E18.5)
MP:0011100	complete preweaning lethality	death of all organisms of a given genotype in a population between fertilization and weaning age (Mus: approximately 3-4 weeks of age)
MP:0011101	partial prenatal lethality	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between fertilization and birth (Mus: approximately E18.5)
MP:0011102	partial embryonic lethality	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms within the embryonic period prior to organogenesis (Mus: prior to E14)
MP:0011103	partial embryonic lethality at implantation	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms at the point of implantation (Mus: E4.5)
MP:0011104	partial embryonic lethality before implantation	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between fertilization and implantation (Mus: E0 to less than E4.5)
MP:0011105	partial embryonic lethality between implantation and placentation	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between the point of implantation and the initiation of placentation (Mus: E4.5 to less than E9)

MP:0011106	partial embryonic lethality between implantation and somite formation	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between the point of implantation and somite formation (Mus: E4.5 to less than E8)
MP:0011107	partial embryonic lethality between somite formation and embryo turning	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between somite formation and the initiation of embryo turning (Mus: E8 to less than E9)
MP:0011108	partial embryonic lethality during organogenesis	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between embryo turning and the completion of organogenesis (Mus: E9-9.5 to less than E14)
MP:0011109	partial lethality throughout fetal growth and development	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between the completion of organogenesis and birth (Mus: E14 to approximately E18.5)
MP:0011110	partial preweaning lethality	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between fertilization and weaning age (Mus: approximately 3-4 weeks of age)
MP:0011111	complete lethality during fetal growth through weaning	death of all organisms of a given genotype in a population between the completion of organogenesis and weaning age (Mus: E14 to approximately 3 weeks of age)
MP:0011112	partial lethality during fetal growth through weaning	the appearance of lower than Mendelian ratios of organisms of a given genotype due to death of some, but not all of the organisms between the completion of organogenesis and weaning age (Mus: E14 to approximately 3 weeks of age)
MP:0011400	complete lethality	all individuals of a given genotype in a population die before the end of the normal lifespan but time(s) of death are unspecified
MP:0013292	embryonic lethality prior to organogenesis	death prior to the completion of embryo turning (Mus: E9-9.5)
MP:0013293	embryonic lethality prior to tooth bud stage	death prior to the appearance of tooth buds (Mus: E12-E12.5)
MP:0013294	prenatal lethality prior to heart atrial septation	death prior to the completion of heart atrial septation (Mus: E14.5-15.5)

E, embryonic day; Mus, Mus musculus.

Table 2.2 Human orthologs of mutation-intolerant essential genes that are not currently included in the catalogs of Mendelian disease genes.

Gene	Chrom	Start	End	RVIS percentile	pLI percentile
ATP6V1B2	8	20197367	20226819	21.6	9.6
CASC3	17	40140318	40172183	9.1	17.1
CASZ1	1	10636604	10796650	13.6	3.4
CHD4	12	6570083	6607476	0.7	0.6
COG3	13	45464898	45536630	17.9	5.7
CTR9	11	10750987	10779743	6.5	1.5
DCP2	5	112976702	113020970	22.0	19.7
DHX30	3	47802909	47850195	2.0	1.5
FOXJ3	1	42176539	42335877	21.3	12.1
MAT2A	2	85539165	85545280	19.3	23.7
MYO18A	17	29073517	29180412	12.8	5.3
NEMF	14	49782083	49853203	5.7	11.5
PRDM10	11	129899706	130002835	15.6	6.9
PREP	6	105277565	105403084	5.9	9.7
RSBN1	1	113761832	113812476	13.0	21.4
RUFY3	4	70704204	70807315	23.1	11.6
SMC5	9	70258962	70354888	9.7	11.7
SYNPO2	4	118850688	119061247	6.8	23.5
TMEM63B	6	44126914	44155519	5.5	4.1
ZMIZ2	7	44748581	44769881	4.9	4.9
ZNF496	1	247297412	247331846	19.9	11.2
ZNF536	19	30228290	30713538	1.1	13.5

Table 2.3 Essentiality status of human disease genes categorized by phenotypic abnormality.

HPO ID	Name	# Genes	# EGs	# NEGs	# Unknown	OR	OR1 95%CI low	OR 95% CI high	P value
HP:0001197	Abnormality of prenatal development or birth	390	234	83	73	3.70	2.87	4.78	3.37E-27
HP:0000769	Abnormality of the breast	230	135	58	37	2.99	2.19	4.08	4.16E-13
HP:0003011	Abnormality of the musculature	1624	798	410	416	2.82	2.48	3.20	4.41E-60
HP:0040064	Abnormality of limbs	1194	639	327	228	2.74	2.38	3.15	2.38E-47
HP:0001608	Abnormality of the voice	215	114	55	46	2.65	1.92	3.67	1.47E-09
HP:0001507	Growth abnormality	1470	735	397	338	2.63	2.31	3.00	2.18E-50
HP:0002086	Abnormality of the respiratory system	1016	521	272	223	2.62	2.25	3.06	6.78E-37
HP:0000924	Abnormality of the skeletal system	1839	909	528	402	2.51	2.24	2.83	7.10E-56
HP:0000152	Abnormality of head or neck	1875	908	539	428	2.45	2.18	2.76	1.64E-53
HP:0025031	Abnormality of the digestive system	1542	760	443	339	2.43	2.15	2.76	2.32E-45
HP:0001626	Abnormality of the cardiovascular system	1418	700	422	296	2.32	2.04	2.64	1.36E-38
HP:0000478	Abnormality of the eye	1755	825	522	408	2.25	2.00	2.53	8.36E-42
HP:0003549	Abnormality of connective tissue	867	432	259	176	2.23	1.90	2.62	1.65E-23
HP:0000119	Abnormality of the genitourinary system	1435	687	429	319	2.23	1.96	2.53	4.39E-35
HP:0002664	Neoplasm	557	303	179	75	2.22	1.84	2.69	4.93E-17

HP:0000707	Abnormality of the nervous system	2336	1064	713	559	2.20	1.98	2.45	5.32E-49
HP:0000598	Abnormality of the ear	1273	605	380	288	2.18	1.91	2.50	3.07E-30
HP:0001574	Abnormality of the integument	1555	731	498	326	2.04	1.80	2.30	1.44E-30
HP:0002715	Abnormality of the immune system	1083	503	365	215	1.84	1.60	2.12	2.55E-17
HP:0001871	Abnormality of blood and blood-forming tissues	827	385	279	163	1.81	1.55	2.13	2.62E-13
HP:0000818	Abnormality of the endocrine system	798	386	281	131	1.81	1.54	2.12	3.03E-13
HP:0001939	Abnormality of metabolism/homeostasis	1500	647	507	346	1.72	1.52	1.95	9.49E-18
HP:0025142	Constitutional symptom	394	173	156	65	1.41	1.13	1.76	0.0022
HP:0045027	Abnormality of the thoracic cavity	18	7	8	3	1.10	0.40	3.03	1

Table 2.4 Tissue-specific essential genes.

Tissue	# Genes	Genes
Testis	27	*LIN28A, DMBX1, FSHR, TRIM71, DPPA4, ZBTB20, NKX1-1, GCM2, GCM1, PKD1L1, PAX4, C7orf55, FGF8, UTF1, SP7, FOXN4, RAD9B, POU4F1, HBZ, TEX19, SKOR2, DCC, C19orf67, MED26, C21orf59, TBX22, PLAC1*
Liver	15	*SERPINC1, APOB, PROC, SLC2A2, FGG, CYP7A1, SAA2, F2, CPB2, F7, CYP1A2, HP, APOH, SERPIND1, F9*
Muscle	12	*AMPD1 ,CACNA1S, MYOG, NEB, CHRNA1, MYL1, CHRNG, TAL2, MYF6, MYF5, ATP2A1, MYLPF*
Kidney	8	*NPHS2, SLC34A1, KCNJ1, GDF3, AQP2, SLC12A1, WNT9B, DNMT3L*
Brain	7	*OTP,CALCR,CHAT,FGF3,GSX1,RTL1,AVP*
Pituitary	6	*LMX1A,POMC,PROP1,LHX3,NEUROD4,RAX*
Adrenal Gland	6	*STAR,CYP11B2,CYP17A1,PHOX2A,CYP11A1,MC2R*
Pancreas	5	*SPINK1,CLPS,PTF1A,IFITM5,INS*
Cells	5	*CR1L,COL10A1,T,HMX3,PPAN*
Whole Blood	4	*F11R,HBB,KLF1,ALAS2*
Thyroid	3	*F11R,FOXE1,TSHR*
Small Intestine	3	*FGF19,MEP1B,P2RY4*
Heart	3	*BMP10,MYL7,MYH6*
Lung	3	*SFTPB,CSF2,SFTPA1*
Nerve	2	*TMEM8C,FGF4*
Skin	2	*HELT,KRT2*
Stomach	1	*RFX6*
Spleen	1	*SPIC*
Bladder	1	*UPK2*
Colon	1	*SLC26A3*
Prostate	1	*SP8*
Esophagus	1	*ERVFRD-1*

Supplementary data

Supplementary data 2.1 IMPC subviable genes with disease causing mutations in HGMD.

Supplementary data 2.2 IMPC lethal genes with disease causing mutations in HGMD.

Supplementary data 2.3 Catalog of EGs and NEGs.

Supplementary data 2.4 Enrichment of EGs among genes within each cytoband in human genome build hg19.

Supplementary data 2.5 Enrichment of EGs among genes within each cytoband in mouse genome build mm10.

CHAPTER 3: Cumulative effect of deleterious variants in essential genes on ASD risk

Introduction

Autism spectrum disorder (ASD) is a heterogeneous, heritable neurodevelopmental syndrome characterized by impaired social interaction, communication, and repetitive behavior (Huguet et al., 2013; State and Levitt, 2011). The highly polygenic nature of ASD (de la Torre-Ubieta et al., 2016; De Rubeis and Buxbaum, 2015; Willsey and State, 2015) suggests that the analysis of the full spectrum of sequence variants in hundreds of genes will be necessary for deeper understanding of disrupted neuronal function. Prioritization of ASD risk genes initially focused on known pathways with recognized relevance to pathogenesis of ASD, such as synaptic function and neuronal development (Geschwind and Levitt, 2007). However, combined analyses of *de novo*, inherited, and case–control variation in over 2,500 ASD parent–child nuclear families identified around 100 genes contributing to ASD risk (De Rubeis et al., 2014; Iossifov et al., 2014; Sanders et al., 2015), converging on pathways implicated in transcriptional regulation and chromatin modeling in addition to synaptic function. The early on-set of ASD suggests a prenatal origin of ASD. Multiple lines of evidence implicated that impairments of early brain development were involved in the pathogenesis of ASD (Parikshak et al., 2013; Stoner et al., 2014; Willsey et al., 2013a).

The main challenge in the current understanding of genetic architecture of ASD comes from a need to study the interplay between variants with a high effect (for example, recurrent *de novo* variants) and a background of variants with an intermediate effect but nevertheless, which still disrupt proper neuronal development. Essential genes (EGs) or genes that are necessary for successful completion of pre- and postnatal development are prime candidates for the source of this background or load of variants with a cumulative intermediate effect. EGs are highly enriched for human disease genes and under strong purifying selection (Georgi et al., 2013; Petrovski et al., 2013; Zhang et al., 2011). In addition to intolerance to loss-of-function , the functional impact of EGs is reflected by haploinsufficiency that is commonly observed in heterozygous mutations (Deutschbauer et al., 2005; Georgi et al., 2013). In addition to their role in defining a "minimal gene set" (Koonin, 2003; Mushegian and Koonin, 1996), EGs tend to play important roles in protein interaction networks (Hwang et al., 2009). Therefore, one may consider that EGs are involved in rate-limiting steps that affect a range of disease pathways (Chakravarti and Turner, 2016).

A deeper understanding of the mutational spectrum of EGs in a neurodevelopmental disorder, such as ASD, is important, because EGs are less likely to be redundant, are more likely to have functional consequences when mutated, and may produce a gradation of phenotypes (White et al., 2013). Previous work by Georgi et al. reported an enrichment of EGs among genes with *de novo* mutations in ASD patients (Georgi et al., 2013). Several groups reported an enrichment of *de novo* and rare inherited single-nucleotide loss-of-function variants in ASD probands (Iossifov et al., 2014; Krumm et al., 2015),

although there is a depletion of damaging mutations in ASD risk genes in population controls (Iossifov et al., 2015; Petrovski et al., 2013; Samocha et al., 2014). With the most comprehensive list of human EGs to our knowledge, I extended the analysis to both *de novo* and inherited damaging variants in 1,781 ASD families. In addition to disease status, I further showed the effect of damaging variants in EGs on ASD-related traits, such as the social skill measurement in 2,348 ASD probands. Finally, I performed coexpression analysis of EGs in the developing human brain to identify clusters of interacting EGs that contribute to ASD risk and suggest ASD candidate genes.

Results

Increased burden of deleterious mutations in essential genes in ASD probands

To address a possible cumulative effect of variants in EGs in ASD in a larger cohort of 1,781 ASD quartet families (with 1,781 probands and 1,781 siblings) from the Simons Simplex Collection (Fischbach and Lord, 2010), I acquired *de novo* and rare inherited mutations from the exome sequencing data of these families (Iossifov et al., 2014; Krumm et al., 2015). I examined the individual mutational burden defined by the number of *de novo* loss-of-function (dnLoF), de novo nonsynonymous damaging (dnNSD), and inherited rare damaging (inhRD) mutations per individual (Supplementary data 3.1 and 3.2). On average, an ASD proband carried 0.06 dnLoF, 0.21 dnNSD, and 10.74 inhRD mutations in EGs. The mutational burden in EGs was significantly elevated in ASD probands compared with unaffected siblings for the three classes of variants considered (P value = 4.75×10^{-7} for dnLoF, P value = 3.41×10^{-4} for dnNSD, and P value = 0.017 for inhRD; one-sided Wilcoxon signed ranked test) (Figure 3.1 and Table 3.1). In

contrast, no significant difference in mutational burden in NEGs was observed (P value = 0.10 for dnLoF, P value = 0.069 for dnNSD, and P value = 0.75 for inhRD) (Table 3.1). Interestingly, 10,823 genes that are currently not assigned as EG or NEG (i.e., phenotypically uncharacterized in mouse knockouts and human cell-based assays) have a moderately elevated burden of dnLoF but not dnNSD and inhRD variants in ASD probands (P value = 0.0042) (Table 3.1). Notably, the effect sizes of EG burden in each variant type correspond to our understanding of the severity of the variant type; *de novo* mutations, which are expected to have a larger functional impact, also display the strongest difference between ASD probands and unaffected siblings (effect size = 0.117 for dnLoF; effect size = 0.079 for dnNSD; Cohen's d). In contrast, inherited mutations are expected to have a moderate functional impact, and a smaller difference is observed between probands and siblings (effect size = 0.042 for inhRD). Although I observed marginally increased burden of dnLoF and dnNSD mutations in EGs in female (n = 325) compared with male (n = 2,043) probands (Table 3.2), the analysis of families divided by gender of proband–sibling pairs (female–female, male–female, female–male, and male–male) showed that gender bias does not underlie the observed differences in mutational burden between probands and siblings (Table 3.3).

The effect of rare damaging mutations in essential genes on social and cognitive impairments

To evaluate the effect of rare damaging mutations in EGs on ASD-associated traits, we used the available quantitative phenotype data on social and cognitive impairments in ~2,500 ASD families from Simons Simplex Collection (Iossifov et al., 2014; Krumm et

al., 2015) (Supplementary data 3.3). As a measure of sociability, I used the total raw score from the Social Responsiveness Scale (SRS) (Constantino and Gruber, 2005), and as cognitive measures, I used three different intelligence quotient (IQ) scores (full-scale IQ, verbal IQ, and nonverbal IQ). As previously reported (Constantino et al., 2003), SRS scores were unrelated to IQ, especially in subjects with IQ higher than 50 (Figure 3.2). In male probands, I observed that the mutational burden in EGs was positively correlated with the SRS total raw score (P value = 1.08×10^{-6}; Poisson regression) (Table 3.4). The effect was not significant in NEGs (P value = 0.21). In female probands, mutational burden in NEGs but not EGs was negatively correlated with SRS total raw score (P value = 0.085 for EG and P value = 6.06×10^{-6} for NEG). In addition, I found that mutational burden in both EGs and NEGs had a significant effect (P value < 2.2×10^{-16}) on verbal and nonverbal IQ scores and that the effect sizes of mutational burden in EGs and NEGs were comparable (Table 3.5). These results suggest that, in ASD probands, deleterious variants in EGs contribute to decreased social skills in males, whereas deleterious variants in both EGs and NEGs lead to decreased IQ.

The overlap between essential genes and known ASD risk genes

To initially explore the overlap between EGs and known ASD genes, I examined the essentiality status of ~500 ASD candidate genes from the Simons Foundation Autism Research Initiative (SFARI) AutDB database (updated December of 2015) (Abrahams et al., 2013) (Figure 3.3). Compared with NEGs, EGs were enriched among ASD candidates categorized as “syndromic” (category S: odds ratio = 3.95, P value = 0.0003; two-sided Fisher’s exact test), candidates with “high confidence” (category 1: odds ratio

= 15.12, P value = 0.0004), and candidates with "suggestive evidence" (category 3: odds ratio = 2.14, P value = 0.0006). Trends of enrichment of EGs were also observed for "strong candidates" (category 2: odds ratio = 1.62, P value = 0.21). I did not observe enrichment of EGs among candidate genes with less supportive evidence (categories 4–6).

To further address whether EGs contribute to ASD risk, I compared the strength of ASD association signals between EGs and NEGs in data from a recent comprehensive analysis of ASD genomic architecture (Sanders et al., 2015), where the transmission and *de novo* association (TADA) test (He et al., 2013) was used to evaluate ASD association based on combined evidence from *de novo* single nucleotide variants (SNVs), *de novo* small deletions, and rare inherited variants from Simons Simplex Collection cohorts as well as case–control data from Autism Sequencing Consortium (ASC) cohorts (Buxbaum et al., 2012). There was a significant enrichment of EGs compared with NEGs in 65 high-confidence TADA ASD genes [TADA false discovery rate (FDR) q values < 0.1] identified by Sanders et al. (Sanders et al., 2015) (36 EGs vs. 15 NEGs; odds ratio = 3.03, P value = 1.82×10^{-4}; one-sided Fisher's exact test). In a broader set of 441 "potential" TADA ASD genes (TADA FDR < 0.5), EGs were also enriched compared with NEGs (132 EGs vs. 117 NEGs; odds ratio = 1.43, P value = 0.00537). Furthermore, by comparing the observed TADA FDR with the expected TADA FDR, I detected a strong deviation from the null distribution in EGs, especially in 132 EGs with potential ASD association (TADA FDR < 0.5) (Figure 3.4). In contrast, NEGs were not enriched for association relative to the background expectation, suggesting that the association signals

between EGs and ASD were stronger and less likely to be false positive compared with NEGs.

The spatio-temporal expression specificity of essential genes in human brain

To explore the spatio-temporal expression specificity of EGs in human brain, I performed Cell type-Specific Expression Analysis (CSEA) (Xu et al., 2014) of EGs and NEGs based on BrainSpan RNA-seq data across 6 brain regions and 10 developmental stages and ages (Table 3.6). Strikingly, I observed distinct temporal patterns of expression specificity between EGs and NEGs (Figure 3.6). In six brain tissue types (amygdala, cerebellum, cortex, hippocampus, striatum and thalamus), EGs were enriched for genes specifically expressed in brain at early developmental stages (from early to mid-fetal) while NEGs were enriched for genes specifically expressed in brain at later stages (from late fetal to young adulthood). For each set of EGs which were specifically expressed, I evaluated the enrichment of 441 potential ASD genes (FDR < 0.5) from Sanders et al. (Sanders et al., 2015) (Figure 3.7; Table 3.6). Significant enrichment was found in early mid fetal cortex (p-value=2.17×10^{-6}; One-sided Fisher's exact test) and early mid fetal striatum (p-value=1.77×10^{-4}; One-sided Fisher's exact test). This finding agrees with recent studies using different approaches which suggested key brain regions involved in the pathogenesis of ASD during early to mid-fetal development (Parikshak et al., 2013; Willsey et al., 2013b; Xu et al., 2014). Moderate enrichment was also found in early fetal cortex, early fetal cerebellum and late mid fetal cortex. In contrast, specifically expressed NEGs showed no enrichment.

Coexpression modules in the developing human brain

It is our hypothesis that a cumulative effect of deleterious variants in several EGs, within the same pathway or across pathways may underlie impaired brain development and individual's ASD risk. To identify clusters of potentially interacting genes, I evaluated the spatiotemporal expression of EGs and NEGs using RNA sequencing (RNA-seq) data from BrainSpan . I identified 41 coexpression modules with distinct expression patterns across 16 brain regions and 31 pre- and postnatal time points (Figure 3.8). I observed that the majority of EG-enriched modules (11 of 14; FDR < 0.1; two-sided Fisher's exact test) (Figure 3.8 and 3.9; Table 3.7) exhibited an "early-expression" pattern, where the expression levels were higher at early fetal stages (starting from 8 postconceptual weeks) and gradually declined before birth. In contrast, the majority of the NEG-enriched modules (15 of 18) exhibited a "later-expression" pattern, with expression levels that were lower at early fetal stages and gradually increased until birth.

I found that EGs in three EG-enriched modules (M01, M02, and M16) were significantly enriched (FDR < 0.1; one-sided Fisher's exact test) for 441 potential TADA ASD genes (Figure 3.9). Notably, all of the three modules were also EG-enriched and early-expressed across fetal brain regions (Figure 3.9 and 3.10). From the pathway enrichment analysis of these EG-enriched modules in the Reactome database (Croft et al., 2014; Fabregat et al., 2016), I found that the top pathways enriched included "transcription" (M01), "chromatin modifying enzymes and chromatin organization" (M02), and "axon guidance" (M16) (Table 3.8), in agreement with the insights from recent large-scale autism studies showing that genes for synaptic formation, transcriptional regulation, and

chromatin remodeling are disrupted in autism (De Rubeis et al., 2014; Iossifov et al., 2014; Sanders et al., 2015). This combined analysis identified 974 EGs from three modules that are coexpressed with known ASD candidate genes at distinct stages of brain development.

29 essential genes as strong candidates for ASD

To further prioritize known EGs as candidates for ASD, I constructed a coexpression network for 974 EGs from three modules enriched for potential ASD genes (Figure 3.10); 844 genes among 974 have a close interaction with high-confidence ASD genes (connected to at least two genes with TADA FDR < 0.1), and 370 genes harbor *de novo* or inherited loss-of-function mutations in ASD individuals from Simons Simplex Collection or ASC cohorts. Of these, 52 have a TADA FDR less than 0.5. Among 52 genes, 23 have been previously shown to contribute to ASD risk [categories syndromic (S), 1, 2, 3, and 4 in SFARI]. For the remaining 29 EGs that have not yet been linked to ASD risk, I argue that, based on (i) the importance of EGs in ASD etiology as shown by their role in critical developmental stages and the increased burden of rare, damaging mutations in ASD individuals; (ii) their coexpression with high-confidence ASD genes in brain; and (iii) the suggestive genetic evidence from the TADA analysis, these 29 EGs represent the strongest candidates for additional investigation in their role in ASD (Figure 3.11 and Table 3.8). According to available mouse phenotypes from the MGI (Eppig et al., 2005) and the IMPC (Koscielny et al., 2014), 11 of these 29 EGs have reported heterozygous phenotypes in mice (Table 3.9). Among them, four EGs (*CHD1*, *FBXO11*,

KDM4B, and *VCP*) have been associated with abnormal neural development and/or behavioral phenotypes in heterozygotes.

Discussion

I provided multiple lines of evidence suggesting that deleterious variants in EGs have a cumulative effect on ASD risk. Using a comprehensive list of 3,915 EGs, I showed that there is both an elevated burden of damaging mutations in EGs in ASD probands and also an enrichment of EGs in the recently identified high-confidence ASD-associated genes. Moreover, the analysis of EGs in the developing brain identified clusters of coexpressed EGs implicated in ASD, including 29 EGs functionally related to previously identified ASD risk genes.

I find that ASD individuals have a higher burden of mutations in EGs compared with their unaffected siblings. It is notable but not surprising that this effect is particularly pronounced when considering *de novo* mutations, because this class of mutations is only subject to selection pressure after originating in the individual, and has exhibited some of the most prominent associations with the risk of ASD (Iossifov et al., 2014; Iossifov et al., 2012; O'Roak et al., 2012b; Sanders et al., 2012) . Similarly, a moderately increased burden of dnLoF variants in ASD probands was detected with a group of 10,823 phenotypically uncharacterized genes. Based on current estimates, one-fifth of these uncharacterized genes (~2,000) are expected to be EGs, which may explain the higher mutational burden of dnLoF variants in ASD probands. Recent studies have begun to show that additional genetic factors, such as rare and common inherited variations, also

contribute to ASD (Gauglerl et al., 2014; Krumm et al., 2015). My results support this finding, showing that inherited, rare, damaging mutations in EGs also have a significant effect on ASD risk. Furthermore, I show an EG-specific effect on social responsiveness, a measure of the social aspects of ASD. In contrast, mutational burden in both EGs and NEGs has an effect on IQ measures. Complex social behaviors result from a range of different cognitive processes; however, in ASD subjects, there is a striking dissociation in the level of impairment in social interaction or communication and general cognitive abilities (as measured by IQ) (Constantino et al., 2003). Moreover, studies in model organisms clearly show a fetal origin for social behavior deficits (Belinson et al., 2016). My results are in line with these findings and suggest that, although a higher mutational burden over all genes may have consequences on IQ, mutational burden in a set of genes with a role at critical early developmental stages influences the development of social behavior. Moreover, my findings are also further supported by the recent report that genomic regions that are under accelerated evolution have essential functions in the human brain development and, when mutated, may cause increased risk for autism (Doan et al., 2016). Therefore, understanding the regulatory landscape of dosage-sensitive EGs expressed at critical stages of brain development may reveal risk alleles for many neurodevelopmental and psychiatric disorders.

The analysis of the overlapping set of Simons Simplex Collection ASD families by several groups using complementary approaches led to the identification of around 100 ASD risk genes and the finding of a depletion of damaging mutations in ASD risk genes (Iossifov et al., 2015; Petrovski et al., 2013; Samocha et al., 2014). I show that a

significant number of reported ASD risk genes are essential for survival and fitness and therefore have a distinctive mutational spectrum, providing a biological foundation for this intolerance to damaging mutations. Of the spectrum of existing alleles, homozygosity or compound heterozygosity for loss-of function alleles will never be observed. Also, because of synthetic lethality, some combinations of mutations in EGs are eliminated. Therefore, individuals will have only a subset of "milder" coding or regulatory alleles. The current list of candidate genes consists of 100 (high-confidence ASD genes) to 400 genes (potential ASD genes) (Sanders et al., 2015). It is striking that my study provides strong statistical evidence for the aggregate effect across 3,915 EGs impacting risk for this neurodevelopmental disorder. A recent SNP-based heritability study reported the extreme polygenicity of schizophrenia, with 70% of 1-Mb genomic regions harboring schizophrenia risk alleles (Loh et al., 2015). Assuming a similar genetic architecture in ASD and schizophrenia, genomic maps of EGs with "surviving" deleterious and regulatory variants in ASD probands represent a complementary approach for the analysis of combinations of culprit genes or alleles.

Because of the fundamental functional role of EGs in an organism, genetic variants in these genes are likely to contribute to many traits and diseases as reflected by the previous finding that EGs are enriched for human disease genes (Dickerson et al., 2011; Dickinson et al., 2016; Georgi et al., 2013). My study is focused on a specific neurodevelopmental disorder—ASD—because it has been suggested that ASD has its roots in abnormalities in prenatal brain development (Hazlett et al., 2017; Parikshak et al., 2013; Stoner et al., 2014; Willsey et al., 2013a). Specifically, my analysis of the

temporal expression patterns of coexpressed gene modules in the developing brain shows that genes in three EG-enriched coexpression modules implicated in ASD are expressed at a high level at the earliest stages of brain development, as early as 8 weeks after conception. In contrast, at later stages of brain development, the expression levels of genes in these EG-enriched modules decrease, whereas the expression levels of genes in NEG-enriched modules increase. The potential role of EGs in ASD is further supported by the analysis of the spatio-temporal expression specificity of essential genes in human brain, which showed that EGs specifically expressed in brain tissues converged at key brain regions (including cortex and striatum) involved in the pathogenesis of ASD (Parikshak et al., 2013; Willsey et al., 2013b; Xu et al., 2014) during early to mid-fetal development. These findings suggest that EGs have a distinctive influence at some of the earliest brain developmental stages as previously reported for constrained genes (Choi et al., 2016) and genes in functional networks perturbed in ASD (Chang et al., 2015). However, to further confirm the distinct contribution of EGs to early onset diseases, a comparison of the burden of deleterious variants in EGs across other complex disorders, including those with a later onset, is warranted.

Each individual can carry a number of deleterious mutations, each of which can have a small effect. Because brain function may be particularly sensitive to mutation accumulation, identifying a specific set of genes in which mutations have a behavioral effect will assist us in understanding how mutation accumulation within an individual can result in a phenotype, such as ASD. Hallmarks of ASD are phenotypic heterogeneity, frequent comorbidities, and that no specific brain region or cell type is uniquely

implicated (de la Torre-Ubieta et al., 2016), further supporting the role of genes with a global effect on embryonic and fetal development. Here, I provide evidence that genes that are essential for survival and fitness also contribute to ASD risk and lead to the disruption of normal social behavior.

Materials and Methods

Burden analysis of mutations in EGs in ASD families

The Simons Simplex Collection contains genetic and phenotypic information from 2,600 ASD families, each of which has one child affected with ASD and unaffected parents and siblings (Fischbach and Lord, 2010). ASD probands were defined by clinical consensus from the Autism Diagnostic Interview–Revised (Lord et al., 1994) and the Autism Diagnostic Observation Schedule (Lord et al., 2000). Multiple individual phenotypic measures, including the Social Responsiveness Scale (SRS) (Constantino and Gruber, 2005) and IQ, were available (Iossifov et al., 2014; Krumm et al., 2015) .

I aimed to investigate the impact of both de novo and rare inherited variants in EGs on ASD risk. I acquired a list of 5,648 de novo variants from an exome sequencing study on 2,517 ASD families from the Simons Simplex Collection (Iossifov et al., 2014) and an additional list of 1,544 *de novo* variants from a reanalysis of the same cohort (2,377 ASD families) with a different pipeline (Krumm et al., 2015). Among 7,192 de novo variants, 674 were loss-of-function mutations (i.e., SNVs that are frameshift, stop-loss, stop-gain, start-loss, splicing donor or acceptor, and frameshift indels), and 3,462 were nonsynonymous mutations (i.e., missense SNVs and nonframeshift indels). The

deleterious *de novo* nonsynonymous mutations were selected using a threshold of the Combined Annotation-Dependent Depletion (CADD) (Kircher et al., 2014) phred-scale score above 10. In addition, I obtained 249,729 rare inherited mutations from 2,377 ASD families (Krumm et al., 2015). From the variants successfully called by both GATK (McKenna et al., 2010) and FreeBayes (Garrison and Marth, 2012), I extracted loss-of-function mutations and nonsynonymous mutations with minor allele frequency in Exome Variant Server (European ancestry) less than 0.01 and CADD phred-scale score above 10. At the end of the variant filtering steps, I obtained 372 dnLoF variants, 1,497 dnNSD variants, and 77,891 inhRD variants in EGs or NEGs for mutational burden analysis (Supplementary data 3.1 and 3.2).

The individual mutational burden was defined as the number of mutations carried by each subject in the gene sets of interest (i.e. 3,915 EGs and 4,919 NEGs) for each class of variants (dnLoF, dnNSD, and inhRD). Among all Simons Simplex Collection ASD families, there were 1,781 ASD quartets where exome sequence data from an affected proband and an unaffected sibling were available. The individual mutational burden in 1,781 ASD probands was compared with the burden in their unaffected sibling using one-sided Wilcoxon signed ranked test. The effect sizes were represented by cohen's D. The 95% confidence interval of effect size was estimated by a bootstrapping procedure, i.e. the 2.5 percentile and 97.5 percentile of the pool of effect sizes from 1000 resampled data sets generated from the original data set. In each resampled data set, the pairs of siblings were randomly selected with replacement.

I acquired SRS total raw scores for 2,348 probands and 1,678 siblings as well as verbal/nonverbal IQs for 2,359 probands for 1,781 ASD quartets and 587 ASD trios from Simons Simplex Collection families (Supplementary data 3.3). Poisson regression analysis was carried out separately between each trait (i.e., SRS total raw score and verbal IQ and nonverbal IQ) as the dependent variables and the individual burdens of all rare damaging mutations (including dnLoF, dnNSD, and inhRD) in EGs or NEGs as the independent variables.

Comparison between observed and expected TADA FDR q values

To compare the strength of association signals to ASD between EGs and NEGs, FDR q values for the TADA test of 18,665 genes were obtained from the work by Sanders et al. (Sanders et al., 2015). For each gene set of interest (i.e., 3,915 EGs or 4,919 NEGs), the null distribution of the transmission and *de novo* association test (TADA) (He et al., 2013) FDR q values was generated by randomly resampling with replacement. Within one iteration of the resampling procedure, the TADA FDR q value of a random gene from the tested 18,665 genes was obtained for each gene in the gene set of interest. The resampled TADA FDR q values were then ranked from low to high. The resampling procedure was repeated for 100,000 iterations. For each observed TADA FDR q value ranked from low to high, the median of 100,000 resampled q values with the same rank was considered the expected TADA FDR q value. The 2.5th and 97.5th percentiles of 100,000 resampled q values were considered the estimated 95% confidence intervals of each expected TADA FDR q value. The observed FDR q values were then compared with the expected FDR q values.

Construction of Coexpression Modules and Coexpression Network in Brain

Coexpression analysis in human brain was conducted based on RNA-seq data from BrainSpan: Atlas of the Developing Human Brain . We used the Weighted Correlation Network Analysis (WGCNA) package (Langfelder and Horvath, 2008) for data quality control and identification of modules of coexpressed genes. The expression data for 52,376 Ensembl genes (Flicek et al., 2014) (including protein-coding genes, noncoding genes, or pseudogenes) across 525 samples were obtained; 1,716 genes with too many missing entries or zero variance in expression levels were removed by the "goodSamplesGenes" function in the WGCNA, and 12,613 genes with very low expression levels [maximum reads per kilobase of transcript per million mapped reads (RPKM) less than 0.5 across samples] were removed. As a final step for gene-level data cleaning, only protein-coding genes were selected for additional analysis. For sample-level data cleaning, three outlier subjects (300, 303, and 306) were removed according to subject-level clustering result. Ten brain tissue types (caudal ganglionic eminence, cerebellum, dorsal thalamus, lateral ganglionic eminence, medial ganglionic eminence, occipital neocortex, parietal neocortex, primary motor sensory cortex, temporal neocortex, and upper rhombic lip) with data from fewer than 10 developmental stages were removed. The final quality-controlled dataset consisted of expression levels of 15,952 protein-coding genes in 16 brain tissue types across 31 pre- and postnatal developmental stages (495 samples in total). For module detection, we used the "blockwiseModules" function in the WGCNA with default parameters, except for the network type (power = 6, deepSplit = 2, and networkType = "signed"). I used the signed version of coexpression

analysis that links two genes with positive correlation of expression levels. Coexpression between gene pairs was calculated based on the quality-controlled BrainSpan RNA-seq data with 495 brain samples. Two genes were defined as “coexpressed” in the brain if the Pearson correlation of the expression levels of both genes across 495 brain samples was greater than or equal to 0.8. In total, there were 8,600,150 coexpression links among protein-coding genes. The coexpression network was created using the GeneMania plugin (Mostafavi et al., 2008) within Cytoscape 3.2.1 (Shannon et al., 2003). Of 974 EGs from three modules (M01, M02, and M16) implicated in ASD, coexpression data were available for 973 genes, which were used as the input gene set for network construction. The coexpression network consists of a main connected component with 963 nodes and 187,443 edges as well as 10 isolated nodes.

Pathway enrichment analysis

I performed pathway enrichment analysis in the Reactome database (Fabregat et al., 2016) using Enrichr (Chen et al., 2013) for three EG-enriched modules (M01, M02, and M16) that were also enriched for potential ASD genes. The enriched pathways were ranked by P values with Benjamini–Hochberg adjustment [False discovery rate (FDR) q values] from the Fisher’s exact test.

The cell type-specific expression analysis (CSEA)

The cell type-specific expression analysis (CSEA) was performed using the SEA on-line tool (http://genetics.wustl.edu/jdlab/csea-tool-2/) (Xu et al., 2014) with lists of 3,915 EGs and 4,919 NEGs as input, among which the SEA analysis were available for 3,838

EGs and 4,757 NEGs. The returned p-values and lists of enriched genes with the threshold of specificity index thresholds (pSI) < 0.05 were used for analysis.

Figures

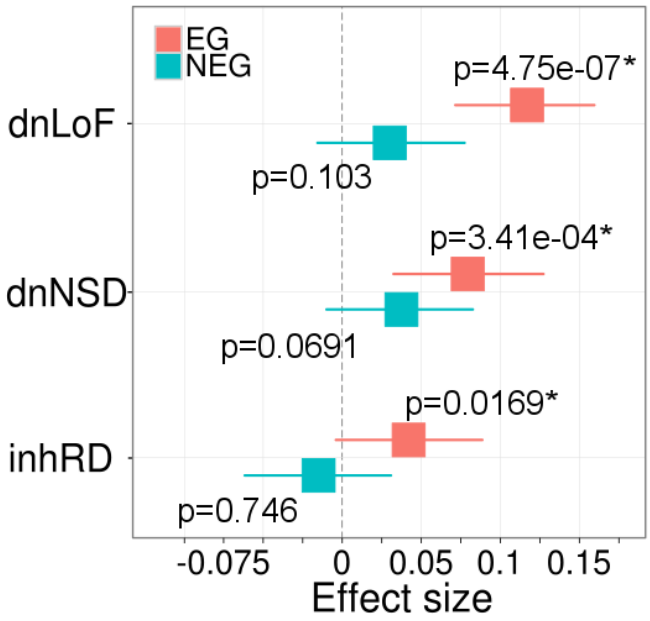

Figure 3.1 Individual mutational burden analysis in 1,781 pairs of ASD probands and unaffected siblings. The analyses were performed separately for 3,915 EGs (red) and 4,919 NEGs (turquoise). The individual mutational burden is defined by the number of dnLoF, dnNSD, and inhRD mutations per individual. Effect sizes were measured by Cohen's d, which is defined as the difference between both means divided by the SD of the paired differences. The estimated 95% confidence intervals of effect sizes were plotted. P values were obtained from one-sided Wilcoxon signed ranked test. *P value < 0.05.

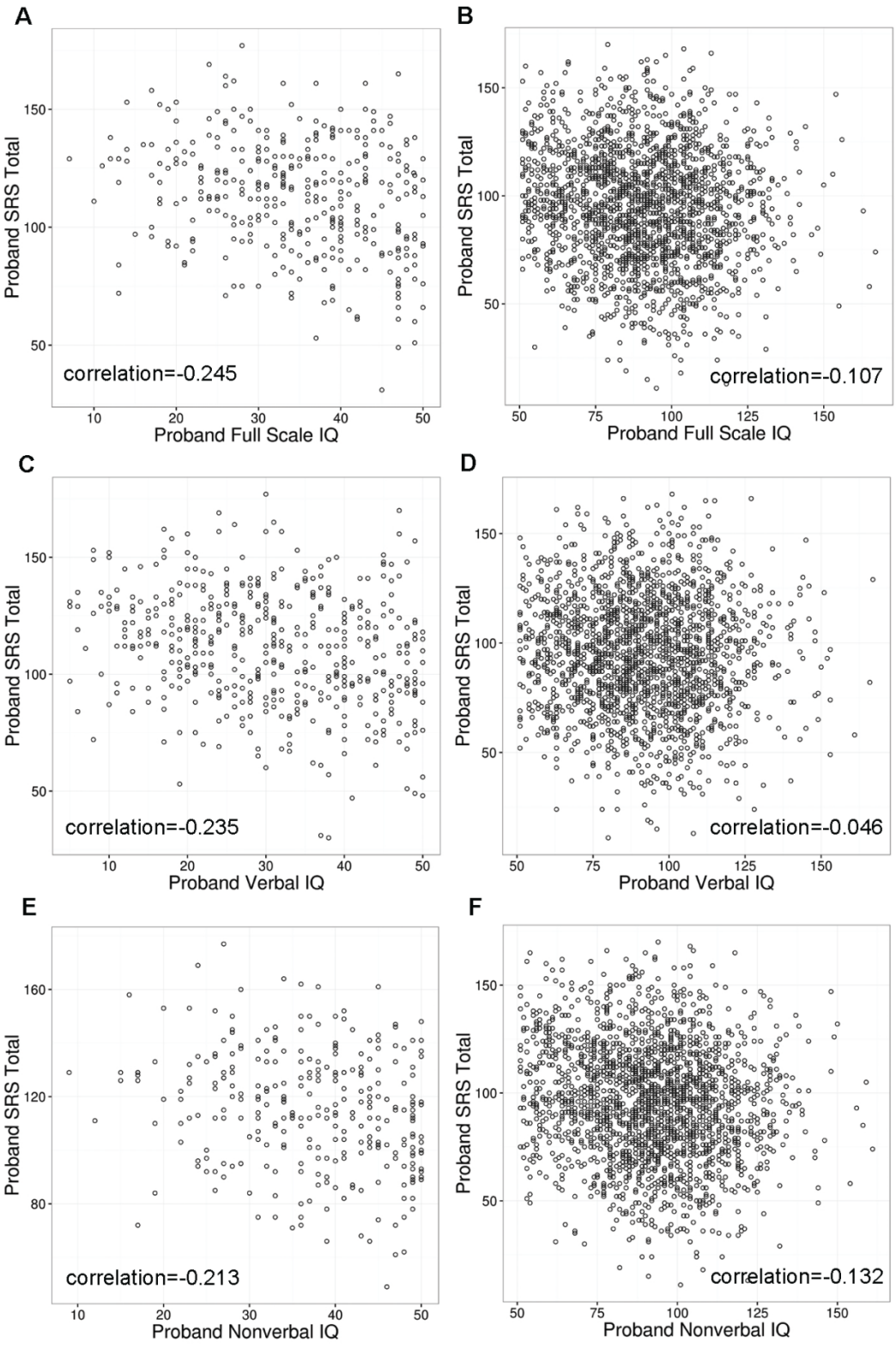
A
Proband SRS Total
Proband Full Scale IQ
correlation=-0.245
B
Proband SRS Total
Proband Full Scale IQ
correlation=-0.107
C
Proband SRS Total
Proband Verbal IQ
correlation=-0.235
D
Proband SRS Total
Proband Verbal IQ
correlation=-0.046
E
Proband SRS Total
Proband Nonverbal IQ
correlation=-0.213
F
Proband SRS Total
Proband Nonverbal IQ
correlation=-0.132

Figure 3.2 Correlation between SRS and IQ. For each of 2,368 ASD probands from Simons Simplex Collection, the Pearson correlation between SRS total raw scores and three IQ scores (full-scale IQ, verbal IQ, and nonverbal IQ) was plotted. The probands were divided by IQ scores: (A, C, and E) IQ< 50 and (B, D, and F) IQ ≥ 50.

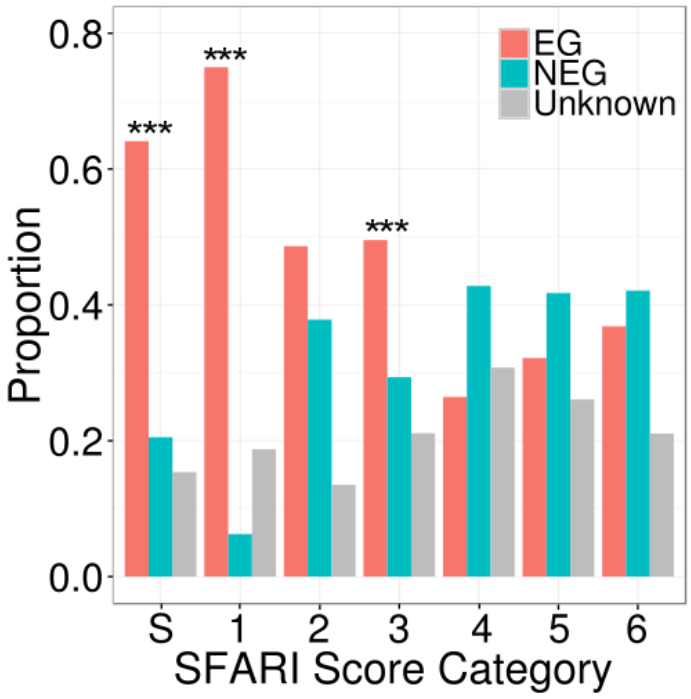

Figure 3.3 Essentiality statuses of SFARI ASD candidate genes. ASD candidate genes categorized by SFARI genes scores [S (syndromic); 1, high confidence; 2, strong candidate; 3, suggestive evidence; 4, minimal evidence; 5, hypothesized; and 6, not supported] (Abrahams et al., 2013) and their essentiality status (EG in red, NEG in turquoise, and unknown in gray). ***The P value from two-sided Fisher's exact test (EG vs. NEG) is less than 0.001.

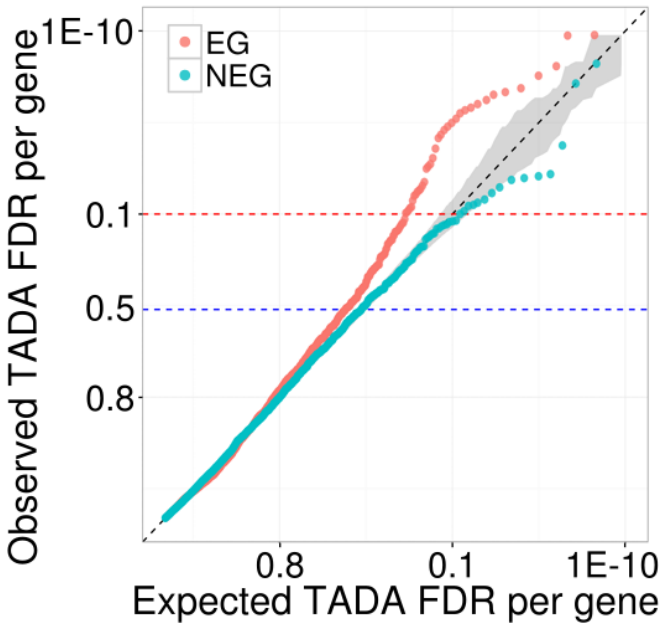

Figure 3.4 The distribution of TADA FDR q values of EGs and NEGs. The FDR q value of the TADA test evaluates ASD association based on combined evidence from de novo SNVs and small deletions, rare inherited variants, and variants (9). The observed negative log10 (q) values of 3,915 EGs (red) and 4,919 NEGs (turquoise) are compared with the expected counterparts under the null hypothesis. The dashed lines indicate the FDR thresholds (FDR = 0.1 in red and FDR = 0.5 in blue) for identification of ASD risk genes. The 95% confidence intervals of the expected negative log10 (q) values are shaded in gray.

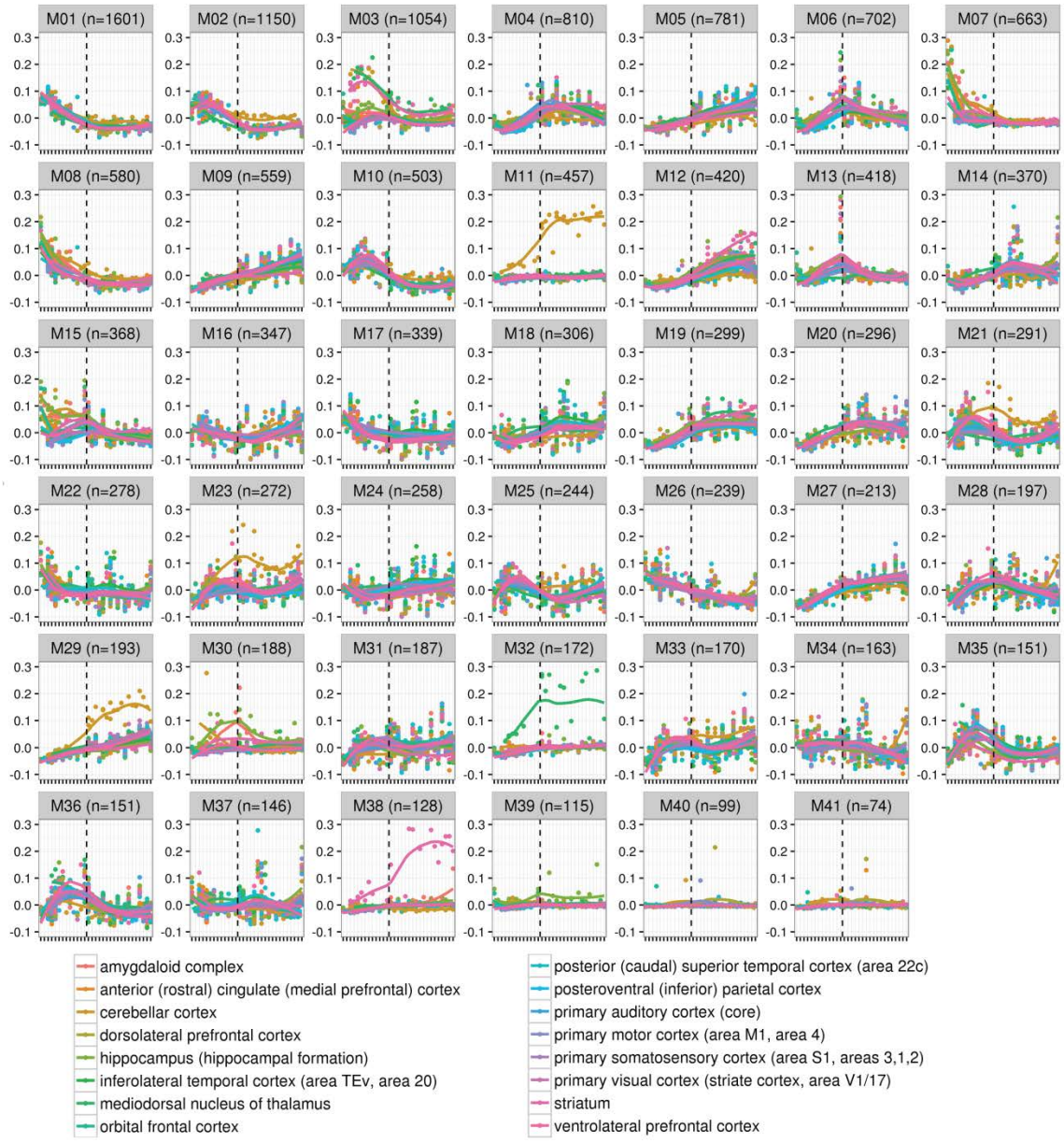

Figure 3.5 Expression profiles of 41 coexpression modules in the brain. Expression profiles of genes from 41 coexpression modules based on the RNA-seq data from BrainSpan are shown. The y axis represents the first principle component of the module-level expression profiles in each brain tissue type. The x axis represents developmental stages in chronological order (Figure 3.7 shows the labels of the time points). The vertical

dashed lines indicate the time of birth. The total number of protein-coding genes in each module (n) is indicated along with the module name.

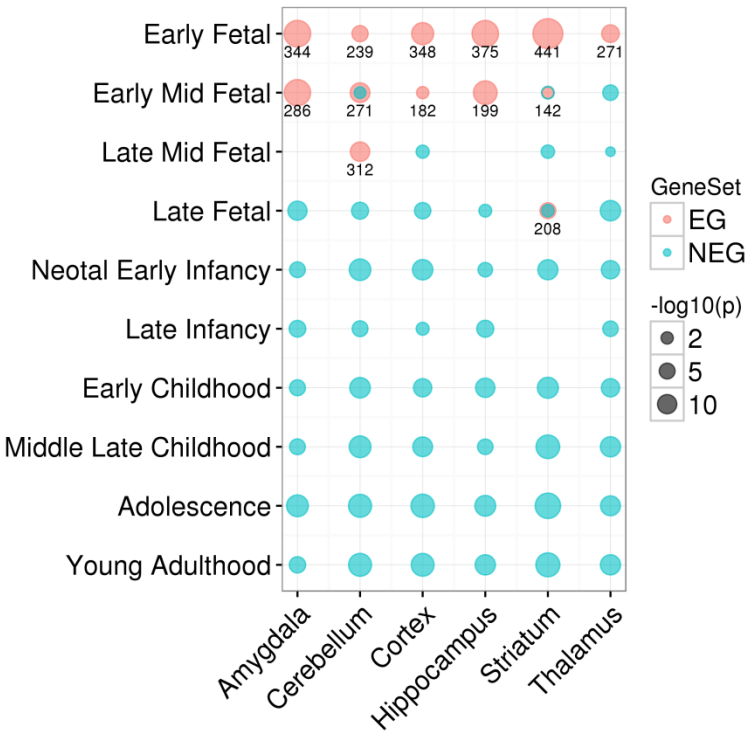

Figure 3.6 Spatio-temporal specific expression of essential genes and non-essential genes. The sizes of dots indicate –log10 of p-values from the specific expression analysis (SEA) on-line tool (http://genetics.wustl.edu/jdlab/csea-tool-2/) (Xu et al., 2014) for 3,915 EGs (in red) and 4,919 NEGs (in turquoise) separately. Each number below red dots indicate the number of specifically expressed EGs in the corresponding brain tissue (x axis) and developmental stage (y axis).

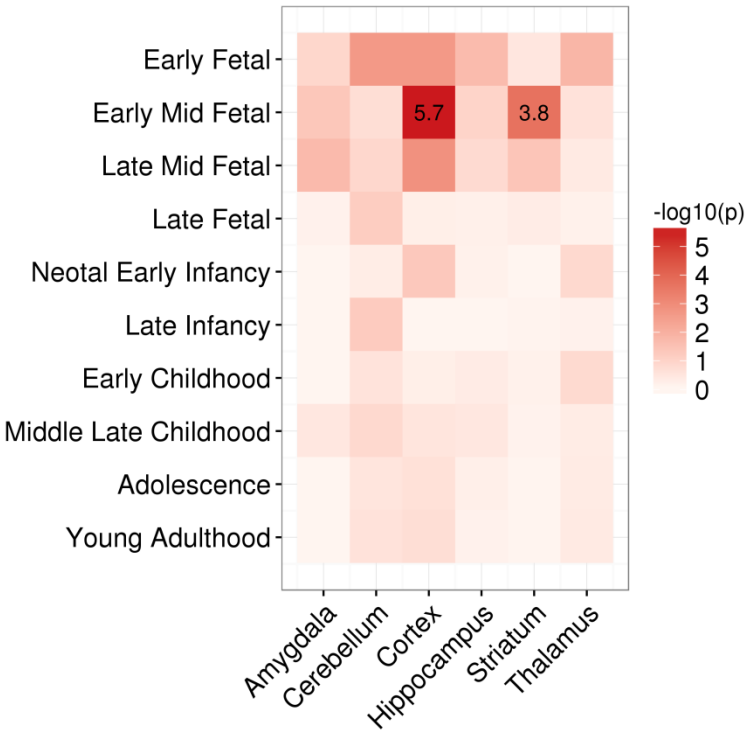

Figure 3.7 Enrichment for potential ASD genes among region- and time-specifically expressed EGs. The heatmap shows the negative log10(p) values for enrichment of 441 potential ASD risk genes (Sanders et al., 2015) in each set of EGs specifically expressed in each brain region (on the x-axis) and developmental stage (on the y-axis). The negative log10 (p) values for combinations of brain regions and developmental stages with significant enrichment of potential ASD genes (p-value<0.05 with Bonferroni correction; one-sided Fisher's exact test) are noted.

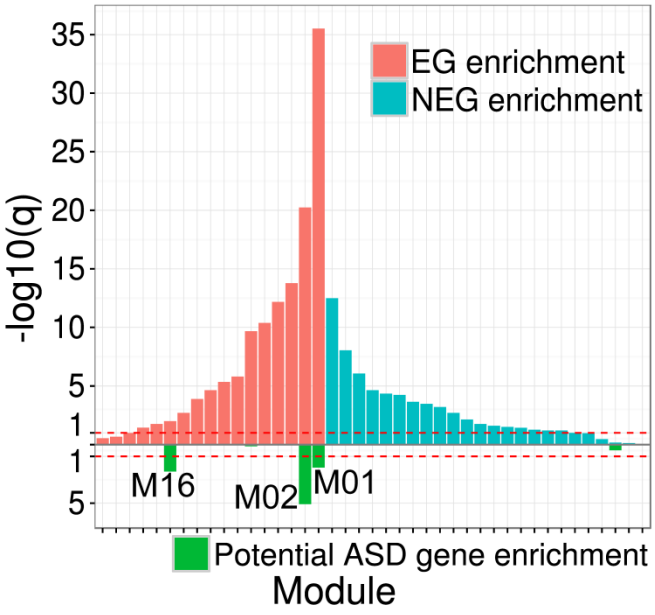

Figure 3.8 Coexpressed modules enriched in EGs and NEGs. The upper barplot displays the level of enrichment of EGs vs. NEGs for each of 41 coexpression modules based on BrainSpan RNA-seq data. The lower barplot displays the level of enrichment (green) of 441 potential ASD genes in EGs from 41 coexpression modules. The heights of the bars represent negative log10 (FDR q value). The upper and lower red dashed lines indicate FDR q value threshold of 0.1.

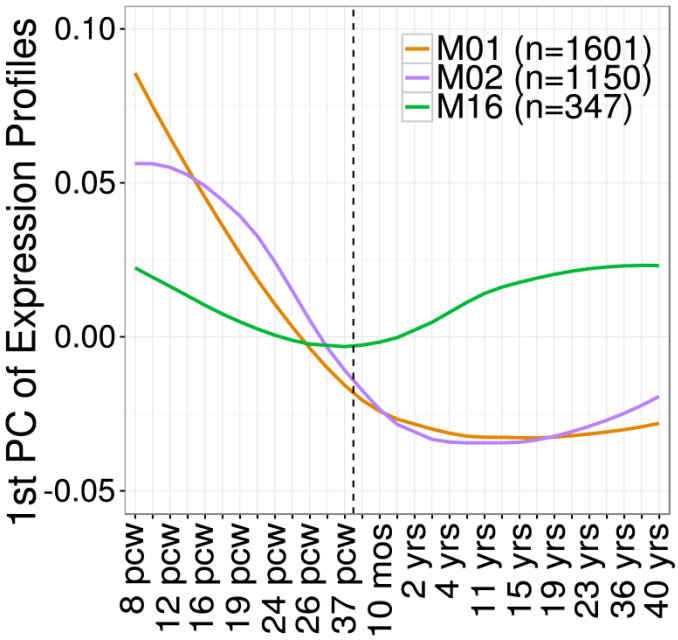

Figure 3.9 The brain expression trajectories of genes from three coexpression modules implicated in ASD. The expression trajectories in brain for 1,601 genes in M01 (orange), 1,150 genes in M02 (purple), and 347 genes in M16 (green) were fitted based on the first principle components of the module level expression profiles (y axis). The x axis represents developmental stages in chronological order. The vertical dashed line indicates the time of birth. pcw, Postconceptual week.

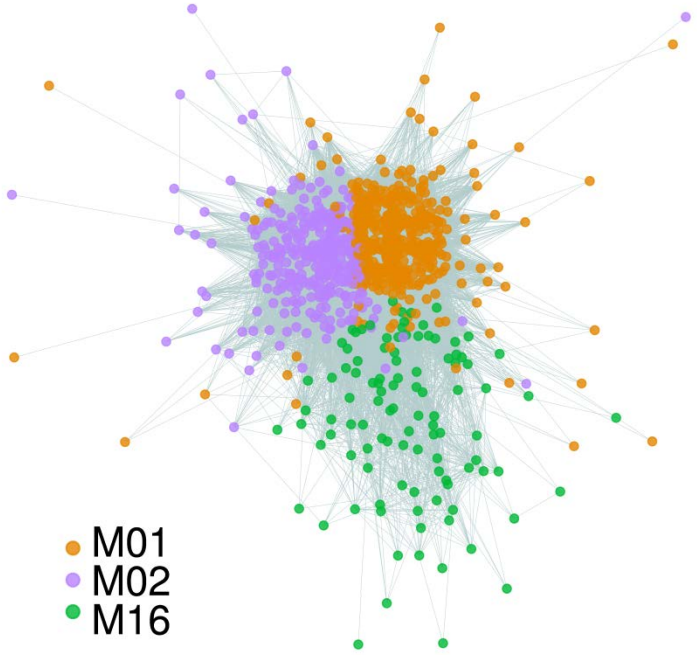

Figure 3.10 Co-expression network of essential genes from three modules implicated in ASD. Coexpression network of 973 of 973 EGs from M01 (orange), M02 (purple), and M16 (green). Edges indicate coexpression between gene pairs.

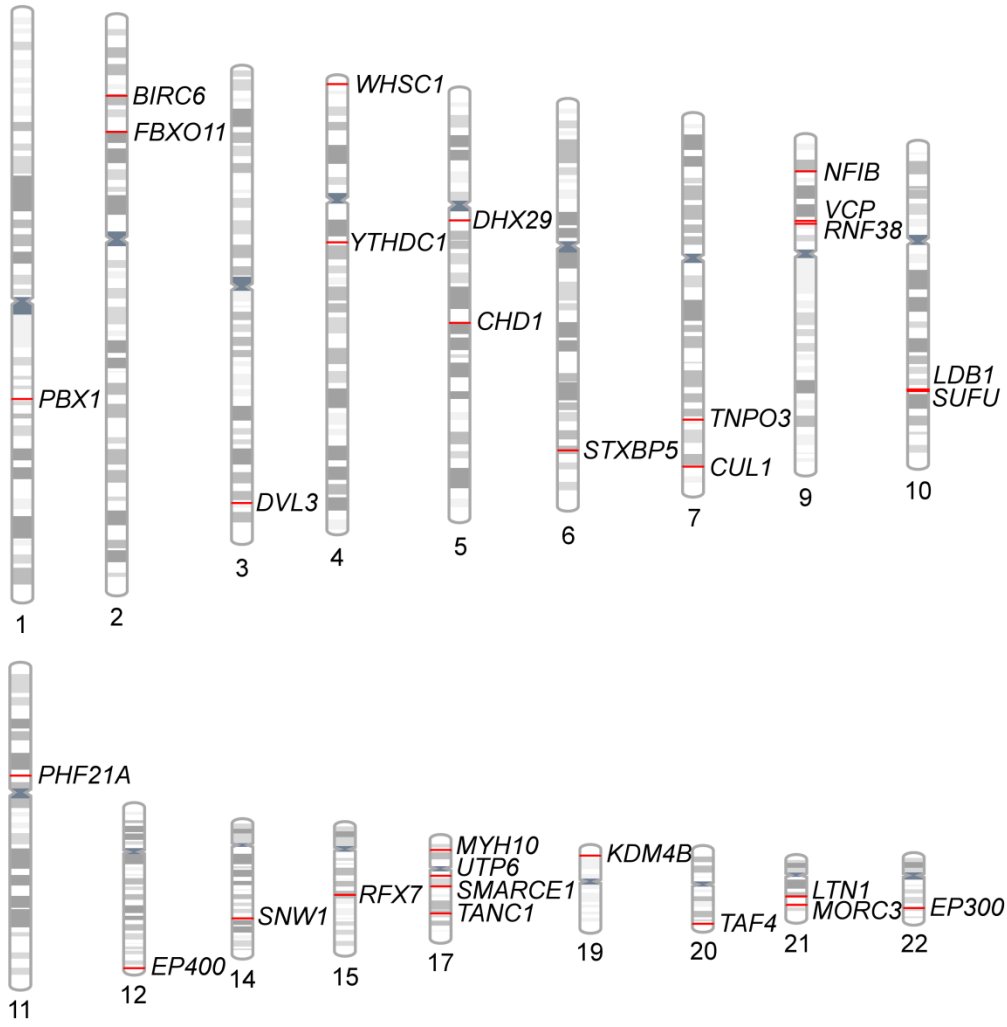

Figure 3.11 Chromosomal distribution of 29 EGs suggested as strong ASD candidate genes. The locations of each gene along the chromosomes are shown in red.

Tables

Table 3.1 Mutational burden analysis in 1,781 ASD quartet families.

Variant Type	Gene Set	# Genes	Proband Average	Sibling Average	Effect Size	Effect Size 95% CI Low	Effect Size 95% CI High	p-value
dnLOF	**EG Ji et al. (current publication)**	3915	0.0640	0.0286	0.1170	0.0715	0.1596	4.75E-07
	EG Dickinson et al. (ref. 14)	3326	0.0595	0.0253	0.1176	0.0730	0.1603	4.16E-07
	EG Georgi et al. (ref. 11)	2472	0.0494	0.0168	0.1254	0.0820	0.1671	7.82E-08
	human cell-EGs (ref. 20, 21, 22)	956	0.0079	0.0056	0.0193	-0.0269	0.0637	0.2118
	NEG	4919	0.0387	0.0309	0.0300	-0.0157	0.0774	0.1028
	Phenotypically uncharacterized genes	10823	0.0752	0.0533	0.0606	0.0143	0.1084	0.004257
dnNSD	**EG Ji et al. (current publication)**	3915	0.2061	0.1589	0.0794	0.0324	0.1274	3.41E-04
	EG Dickinson et al. (ref. 14)	3326	0.1875	0.1376	0.0892	0.0429	0.1353	8.13E-05
	EG Georgi et al. (ref. 11)	2472	0.1505	0.1050	0.0895	0.0435	0.1366	7.36E-05
	human cell-EGs (ref. 20, 21, 22)	956	0.0371	0.0365	0.0021	-0.0435	0.0499	0.4696
	NEG	4919	0.1611	0.1404	0.0374	-0.0100	0.0827	0.0691
	Phenotypically uncharacterized genes	10823	0.2471	0.2791	-0.0419	-0.0884	0.0044	0.9636
inhRD	**EG Ji et al. (current publication)**	3915	10.7428	10.6042	0.0420	-0.0041	0.0887	0.01688
	EG Dickinson et al. (ref. 14)	3326	9.3257	9.2358	0.0287	-0.0185	0.0757	0.04139
	EG Georgi et al. (ref. 11)	2472	7.0236	6.9163	0.0402	-0.0053	0.0867	0.02622
	human cell-EGs (ref. 20, 21, 22)	956	2.3745	2.3779	-0.0022	-0.0485	0.0435	0.5935
	NEG	4919	12.7816	12.8355	-0.0150	-0.0618	0.0308	0.7456
	Phenotypically uncharacterized genes	10823	20.3947	20.4559	-0.0133	-0.0592	0.0342	0.5404

Table 3.2 Difference in individual mutational burden between male and female probands.

Variant Type	Gene Set	Female Proband Average	Male Proband Average	Effect Size	p-value
dnLoF	EG	0.0862	0.0597	0.1042	**0.0355**
	NEG	0.0462	0.0357	0.0551	0.1782
dnNSD	EG	0.2400	0.1948	0.1014	**0.0388**
	NEG	0.2000	0.1596	0.0993	0.0742
inhRD	EG	11.0523	10.9633	0.0151	0.4711
	NEG	13.2677	13.0113	0.0360	0.5271

Effect sizes were measured by Cohen's d, which is defined as the difference between both means divided by pooled standard deviation. P-values with statistical significance are in bold.

Table 3.3 Mutational burden analysis in ASD probands and unaffected siblings (dissected by the genders of proband-sibling pairs).

Variant Type	Gene Set	Proband Gender	Sibling Gender	# Families	Proband Average	Sibling Average	Effect Size	p-value
dnLoF	EG	All	All	1781	0.0640	0.0286	0.1170	4.75×10^{-7}
		Female	Male	101	0.0891	0.0099	0.2588	0.0067
		Male	Female	826	0.0593	0.0327	0.0893	0.0053
		Male	Male	732	0.0615	0.0246	0.1228	0.0005
		Female	Female	122	0.0902	0.0410	0.1461	0.0600
	NEG	All	All	1781	0.0387	0.0309	0.0300	0.1028
		Female	Male	101	0.0396	0.0297	0.0374	0.3884
		Male	Female	826	0.0412	0.0266	0.0558	0.0549
		Male	Male	732	0.0369	0.0314	0.0213	0.2838
		Female	Female	122	0.0328	0.0574	0.0818	0.8302
dnNSD	EG	All	All	1781	0.2061	0.1589	0.0794	0.0003
		Female	Male	101	0.2178	0.1683	0.0724	0.2392
		Male	Female	826	0.2094	0.1755	0.0552	0.0454
		Male	Male	732	0.1885	0.1270	0.1136	0.0013
		Female	Female	122	0.2787	0.2295	0.0725	0.2157
	NEG	All	All	1781	0.1611	0.1404	0.0374	0.0691
		Female	Male	101	0.1881	0.1980	0.0155	0.5696
		Male	Female	826	0.1465	0.1477	0.0022	0.5515
		Male	Male	732	0.1667	0.1175	0.0904	0.0080
		Female	Female	122	0.2049	0.1803	0.0379	0.3817
inhRD	EG	All	All	1781	10.7428	10.6042	0.0420	0.0169
		Female	Male	101	10.3762	10.6436	0.0778	0.8260
		Male	Female	826	10.8341	10.7034	0.0401	0.1120
		Male	Male	732	10.5765	10.4372	0.0417	0.0449
		Female	Female	122	11.4262	10.9016	0.1619	0.0430
	NEG	All	All	1781	12.7816	12.8355	0.0150	0.7456
		Female	Male	101	12.5050	13.0792	0.1398	0.9143
		Male	Female	826	12.8693	13.0182	0.0424	0.7802
		Male	Male	732	12.6134	12.5546	0.0165	0.5576
		Female	Female	122	13.4262	13.0820	0.0907	0.1327

Effect sizes were measured by Cohen's d, which is defined as the difference between both means divided by the standard deviation of the paired differences. P-values were obtained from one-sided Wilcoxon signed rank test.

Table 3.4 Relationship between individual mutational burden and social responsiveness scale in ASD probands.

Group	Gene Set	Estimate	Std. Error	p-value
2031 male probands	EG (3915 genes)	0.001860	0.000381	**1.08×10^{-6}**
	NEG (4919 genes)	0.000407	0.000324	0.209
317 female probands	EG (3915 genes)	-0.001511	0.000877	0.085
	NEG (4919 genes)	-0.003084	0.000682	6.04×10^{-6}

Coefficients for Poisson regression are shown, which modeled the relationship between SRS total raw score and individual burden of all rare damaging mutations (including dnLOF, dnNSD and inhRD mutations). The p-value with statistical significance with positive estimated effects (p-value<0.05, estimate>0) is in bold.

Table 3.5 Relationship between individual mutational burden and IQ in ASD probands.

Trait	Gene Set	Estimate	Std. Error	p-value
Verbal IQ	EG (3915 genes)	-0.007279	0.000400	$<2.2\times10^{-16}$
	NEG (4919 genes)	-0.005307	0.000383	$<2.2\times10^{-16}$
Nonverbal IQ	EG (3915 genes)	-0.007172	0.000336	$<2.2\times10^{-16}$
	NEG (4919 genes)	-0.004906	0.000320	$<2.2\times10^{-16}$

Coefficients for Poisson regression are shown, which modeled the relationship between verbal/nonverbal IQ and individual burden of all rare damaging mutations (including dnLOF, dnNSD and inhRD mutations).

Table 3.6 The spatio-temporal expression specificity of essential genes in human brain.

Tissue	Stage	p-value Genes	# Genes	# pASD	OR pASD	p-value pASD	pASD list
Cortex	Early Mid Fetal	0.01	182	17	4.28	2.17E-06	*NFIB,TRIO,NFIA,ADNP,NUAK1,BCL11A,TBR1,TANC2,GRIN2B,KDM4B,JUP,LDB1,KDM6B,DSCAM,SUV420H1,KDM5B,PBX1*
Striatum	Early Mid Fetal	0.015	142	12	3.80	0.00018	*TRIO,DNMT3A,BCL11A,TANC2,RAI1,LDB1,FOXP1,RELN,KDM6B,SUV420H1,KDM5B,PBX1*
Cortex	Late Mid Fetal	1	109	9	3.68	0.0013	*TRIO,PRPF39,TBR1,SYNGAP1,ANK2,JUP,PTMS,WNT9A,KDM6B*
Cerebellum	Early Fetal	3.77E-06	239	14	2.56	0.0021	*TRIO,FOXP1,ADNP,NR2F1,RFX7,BCL11A,SIX2,PAX5,JUP,DNMT3A,KDM6B,SUV420H1,KDM5B,WNT7B*
Cortex	Early Fetal	1.72E-16	348	18	2.25	0.0022	*TBR1,CHD1,NFIA,PPM1D,TCF3,NFIB,NUAK1,RFX7,BCL11A,MIB1,SIX2,ERI1,TTK,POLD1,KDM6B,TANC2,KDM5B,PBX1*
Thalamus	Early Fetal	4.33E-08	271	13	2.06	0.0151	*CTNNB1,DNMT3A,WHSC1,RFX7,NR2F1,TCF7L2,TTK,JUP,SMARCE1,RNF38,NUAK1,KDM5B,PBX1*
Amygdala	Late Mid Fetal	1	113	7	2.69	0.0200	*TGM1,TRIO,ADNP,NR2F1,PRPF39,TBR1,WNT9A*
Hippocampus	Early Fetal	2.60E-28	375	16	1.83	0.0209	*CHD1,NFIB,NFIA,PPM1D,PTK7,TCF3,WHSC1,NR2F1,TBR1,MIB1,TTK,POLD1,ILF2,ERI1,KDM5B,WNT7B*
Striatum	Late Mid Fetal	1	131	7	2.30	0.0402	*FOXP1,BCL11A,RAI1,ETV2,PTMS,KDM6B,DNMT3A*
Amygdala	Early Mid Fetal	1.13E-27	286	12	1.79	0.0463	*NFIB,DNMT3A,TCF3,WHSC1,RFX7,NR2F1,TBR1,TTK,POLD1,LDB1,KDM6B,PBX1*
Cortex	Neotal Early Infancy	1	58	4	3.01	0.0516	*SYNGAP1,DSG3,SHANK2,PLCB1*
Cerebellum	Late Infancy	1	174	8	1.96	0.0614	*TGM1,PRPF39,PPM1D,INTS6,ETV2,PTMS,RELN,RIMS1*
Cerebellum	Late Fetal	0.231	148	7	2.02	0.0685	*NFIB,NFIA,PTK7,CTNNB1,RELN,OVOL1,WNT7B*
Hippocampus	Early Mid Fetal	2.61E-20	199	8	1.70	0.1110	*NFIB,NFIA,ETV2,PTK7,TCF3,TBR1,TTK,KDM5B*
Amygdala	Early Fetal	1.05E-28	344	12	1.47	0.1320	*PPM1D,ADNP,TCF3,WHSC1,RFX7,NR2F1,TTK,POLD1,DNMT3A,SMARCE1,KDM5B,ILF2*

Cerebellum	Late Mid Fetal	5.54E-11	312	11	1.49	0.1376	*CHD1,NFIB,PTK7,TCF3,TGM1,TTK,EP400,RELN,OVOL1,DSCAM, WNT7B*
Cerebellum	Middle Late Childhood	1	183	7	1.61	0.1564	*TGM1,NFIA,PPM1D,SCN2A,RELN,ATP1A1,RIMS1*
Thalamus	Neotal Early Infancy	1	118	5	1.79	0.1582	*ACHE,WNT9A,RIMS1,TCF7L2,SCN1A*
Thalamus	Early Childhood	1	58	3	2.21	0.1653	*SCN1A,TCF7L2,RGMA*
Hippocampus	Late Mid Fetal	1	89	4	1.91	0.1692	*PLVAP,NR3C2,WNT9A,NFIB*
Cerebellum	Early Mid Fetal	1.15E-11	271	9	1.39	0.2118	*PLVAP,DNMT3A,PTK7,TCF3,JUP,RELN,WNT9A,LAMB1,WNT7B*
Cortex	Young Adulthood	1	100	4	1.69	0.2235	*PLCB1,SCN1A,TBR1,NUAK1*
Cortex	Adolescence	1	73	3	1.73	0.2595	*PLCB1,SCN1A,NUAK1*
Cerebellum	Young Adulthood	1	147	5	1.42	0.2840	*TGM1,SCN2A,RELN,ATP1A1,RIMS1*
Thalamus	Early Mid Fetal	1	113	4	1.48	0.2925	*NR2F1,WNT9A,TCF7L2,RGMA*
Cerebellum	Early Childhood	1	151	5	1.39	0.3028	*TGM1,SCN2A,RELN,ATP1A1,RIMS1*
Cerebellum	Adolescence	1	158	5	1.32	0.3361	*TGM1,SCN2A,RELN,ATP1A1,RIMS1*
Cortex	Middle Late Childhood	1	50	2	1.68	0.3420	*NUAK1,PLCB1*
Striatum	Early Fetal	3.04E-40	441	12	1.13	0.3805	*CHD1,PPM1D,TCF3,WHSC1,RFX7,ERI1,TTK,POLD1,DNMT3A,SM ARCE1,ABL1,PBX1*
Amygdala	Middle Late Childhood	1	21	1	2.02	0.4023	*FERMT1*
Hippocampus	Middle Late Childhood	1	57	2	1.47	0.4028	*NR3C2,NCKAP1*
Thalamus	Young Adulthood	1	105	3	1.19	0.4687	*SHANK3,TCF7L2,SCN1A*
Thalam	Late Mid	1	109	3	1.14	0.4934	*TGM1,WNT9A,TCF7L2*

us	Fetal						
Thalamus	Adolescence	1	113	3	1.10	0.5175	*ACHE,TCF7L2,SCN1A*
Hippocampus	Early Childhood	1	73	2	1.14	0.5304	*NR3C2,MYO1E*
Thalamus	Middle Late Childhood	1	116	3	1.07	0.5351	*FERMT1,TCF7L2,SCN1A*
Striatum	Late Fetal	7.64E-06	208	5	0.99	0.5685	*TTK,POLD1,FERMT1,FOXP1,KDM6B*
Cerebellum	Neotal Early Infancy	1	122	3	1.02	0.5694	*PTK7,RELN,ATP1A1*
Cortex	Late Fetal	1	91	2	0.91	0.6502	*JUP,TBR1*
Cortex	Early Childhood	1	43	1	0.96	0.6516	*GRIN2B*
Hippocampus	Adolescence	1	94	2	0.88	0.6677	*NR3C2,SHANK3*
Hippocampus	Late Fetal	1	105	2	0.78	0.7256	*PLVAP,TTK*
Thalamus	Late Fetal	1	114	2	0.72	0.7665	*ACHE,TCF7L2*
Hippocampus	Neotal Early Infancy	1	60	1	0.68	0.7706	*WNT7B*
Striatum	Early Childhood	1	61	1	0.67	0.7761	*PLCB1*
Thalamus	Late Infancy	1	62	1	0.66	0.7816	*TCF7L2*
Amygdala	Late Fetal	1	119	2	0.69	0.7868	*TBR1,ERI1*
Hippocampus	Young Adulthood	1	64	1	0.64	0.7920	*NR3C2*
Striatum	Middle Late Childhood	1	72	1	0.57	0.8292	*FERMT1*
Striatum	Late Infancy	1	86	1	0.47	0.8790	*RPL12*
Striatum	Young Adulthood	1	107	1	0.38	0.9278	*MYO1E*

Striatum	Adolescence	1	136	1	0.30	0.9647	*PLCB1*
Amygdala	Neotal Early Infancy	1	41	0	0.00	1	
Amygdala	Early Childhood	1	29	0	0.00	1	
Hippocampus	Late Infancy	1	72	0	0.00	1	
Amygdala	Adolescence	1	62	0	0.00	1	
Amygdala	Late Infancy	1	77	0	0.00	1	
Amygdala	Young Adulthood	1	52	0	0.00	1	
Cortex	Late Infancy	1	41	0	0.00	1	
Striatum	Neotal Early Infancy	1	95	0	0.00	1	

pASD, potential ASD genes (n=441); OR, odds ratio. P-values and odds ratios are from one-sided Fisher's exact test s.

Table 3.7 Co-expression modules in the brain.

Module	#Gene	Expression pattern	Enrichment	#EG	#NEG	Odds ratio (EG/NEG)	p-value (EG/NEG)	#Potential ASD genes	Odds ratio (ASD genes)	p-value (ASD genes)
M01	1601	Early expressed	EG-enriched	501	251	2.73	**$7.38*10^{-38}$**	55	1.52	**0.004**
M02	1150	Early expressed	EG-enriched	367	208	2.34	**$2.80*10^{-22}$**	53	2.13	**2.58×10^{-6}**
M03	1054	Mixed	NEG-enriched	204	340	0.74	**$9.67*10^{-4}$**	18	0.72	0.934
M04	810	Late expressed	NEG-enriched	122	326	0.45	**$3.19*10^{-14}$**	19	1.00	0.529
M05	781	Late expressed	NEG-enriched	156	239	0.81	**0.0491**	24	1.32	0.122
M06	702	Late expressed	NEG-enriched	129	254	0.63	**$1.55*10^{-5}$**	11	0.65	0.948
M07	663	Early expressed	EG-enriched	251	141	2.32	**$1.23*10^{-15}$**	8	0.50	0.989
M08	580	Early expressed	EG-enriched	193	114	2.19	**$3.62*10^{-11}$**	13	0.95	0.613
M09	559	Late expressed	NEG-enriched	104	206	0.62	**$9.26*10^{-5}$**	16	1.23	0.246
M10	503	Early expressed	EG-enriched	126	114	1.40	**0.0102**	9	0.74	0.847
M11	457	Late expressed	NEG-enriched	79	178	0.55	**$7.33*10^{-6}$**	9	0.83	0.753
M12	420	Late expressed	NEG-enriched	62	163	0.47	**$1.90*10^{-7}$**	7	0.69	0.874
M13	418	Late expressed	NEG-enriched	97	193	0.62	**$1.46*10^{-4}$**	7	0.69	0.877
M14	370	Late expressed	EG-enriched	81	58	1.77	**0.00102**	4	0.45	0.977

M15	368	Mixed	EG-enriched	104	95	1.39	**0.0251**	5	0.57	0.934
M16	347	Early expressed	EG-enriched	106	90	1.49	**0.00570**	20	2.57	**2.80×10^{-4}**
M17	339	Early expressed	EG-enriched	102	59	2.20	**$1.20*10^{-06}$**	16	2.05	0.008
M18	306	Late expressed		66	61	1.37	0.0874	5	0.67	0.861
M19	299	Late expressed	NEG-enriched	31	118	0.32	**$1.81*10^{-9}$**	2	0.28	0.994
M20	296	Late expressed	NEG-enriched	51	91	0.70	**0.0498**	5	0.72	0.823
M21	291	Early expressed		54	73	0.93	0.719	5	0.72	0.818
M22	278	Early expressed	EG-enriched	83	25	4.24	**$6.17*10^{-12}$**	2	0.29	0.991
M23	272	Late expressed	NEG-enriched	41	84	0.61	**0.0108**	2	0.31	0.988
M24	258	Early expressed		51	49	1.31	0.189	11	1.84	0.047
M25	244	Early expressed	EG-enriched	86	49	2.23	**$6.66*10^{-6}$**	11	1.98	0.031
M26	239	Early expressed	EG-enriched	79	18	5.61	**$8.28*10^{-14}$**	4	0.70	0.821
M27	213	Late expressed	NEG-enriched	45	85	0.66	**0.0261**	6	1.19	0.399
M28	197	Late expressed		32	41	0.98	1	1	0.21	0.991
M29	193	Late expressed	NEG-enriched	33	69	0.60	**0.0158**	2	0.43	0.943
M30	188	Late expressed	NEG-enriched	11	43	0.32	**$2.92*10^{-4}$**	3	0.69	0.808
M31	187	Late expressed		41	64	0.80	0.323	6	1.38	0.279
M32	172	Late expressed	NEG-enriched	24	60	0.50	**0.00388**	3	0.75	0.766

M33	170	Late expressed		41	40	1.29	0.263	4	1.00	0.568
M34	163	Mixed	EG-enriched	48	22	2.76	**$5.06*10^{-5}$**	2	0.51	0.904
M35	151	Mixed	NEG-enriched	21	48	0.55	**0.0207**	6	1.73	0.147
M36	151	Late expressed		22	44	0.63	0.0815	3	0.82	0.707
M37	146	Early expressed	EG-enriched	38	9	5.35	**$3.81*10^{-7}$**	2	0.57	0.862
M38	128	Late expressed	NEG-enriched	17	63	0.34	**$2.11*10^{-5}$**	4	1.36	0.347
M39	115	Early expressed		29	42	0.87	0.632	4	1.47	0.298
M40	99	unknown		4	13	0.39	0.0926	1	0.45	0.890
M41	74	unknown	NEG-enriched	4	16	0.31	**0.0400**	1	0.59	0.816

P-values with statistical significance are in bold.

Table 3.8 Reactome pathways enriched in three EG-enriched modules implicated in ASD.

Module	Term	Overlap	p-value	Adjusted p-value	Genes
M01	Transcription	25/202	$2.40*10^{-6}$	**$6.79*10^{-4}$**	*GTF3C3;HDAC2;CCNT2;GTF3C4;RRN3;CSTF3;GTF2E1;CLP1;PCF11;POLR2B;SNAPC3;CSTF1;RNGTT;TBP;NCBP1;NCBP2;GTF2H3;NFIA;POLR3B;NFIB;POLR3C;POLR1B;POLR1E;TFAM;TAF5*
	Processing of Capped Intron-Containing Pre-mRNA	22/144	$4.34*10^{-7}$	**$3.67*10^{-4}$**	*NCBP1;NUP133;DHX9;NCBP2;CSTF3;CDC5L;HNRNPU;PLRG1;YBX1;NUP160;EFTUD2;PRPF4;CLP1;HNRNPH1;PCF11;POLR2B;NUP50;CSTF1;NUPL1;RAE1;SF3B1;CTNNBL1*
	Folding of actin by CCT/TriC	7/9	$1.11*10^{-6}$	**$4.70*10^{-4}$**	*CCT3;CCT6A;CCT2;TCP1;CCT7;CCT5;CCT4*
	mRNA Splicing	17/113	$1.11*10^{-5}$	**0.00188**	*NCBP1;DHX9;NCBP2;CSTF3;CDC5L;HNRNPU;PLRG1;YBX1;EFTUD2;PRPF4;CLP1;HNRNPH1;PCF11;POLR2B;CSTF1;SF3B1;CTNNBL1*
	HIV Infection	23/218	$6.34*10^{-5}$	**0.00589**	*CCNT2;PSMD11;RNGTT;TBP;TSG101;NCBP1;NUP133;NCBP2;XRCC5;HMGA1;NEDD4L;GTF2H3;GTF2E1;NUP160;AP1G1;POLR2B;NUP50;PSMD2;TAF5;NUPL1;PAK2;RAE1;KPNB1*
	HIV Life Cycle	18/137	$3.19*10^{-5}$	**0.00451**	*CCNT2;RNGTT;TBP;TSG101;NCBP1;NUP133;NCBP2;XRCC5;HMGA1;NEDD4L;GTF2H3;GTF2E1;NUP160;POLR2B;NUP50;TAF5;NUPL1;RAE1*
	snRNP Assembly	10/49	$8.30*10^{-5}$	**0.00589**	*NCBP1;NUP133;NCBP2;NUP50;TGS1;DDX20;NUPL1;RAE1;NUP160;WDR77*
	Formation of tubulin folding intermediates by CCT/TriC	7/20	$5.94*10^{-5}$	**0.00589**	*CCT3;CCT6A;CCT2;TCP1;CCT7;CCT5;CCT4*
	Association of TriC/CCT with target proteins during biosynthesis	8/29	$7.19*10^{-5}$	**0.00589**	*CCT3;CCT6A;CCT2;TCP1;XRN2;CCT7;CCT5;CCT4*
	Regulation of cholesterol biosynthesis by SREBP (SREBF)	10/53	$1.47*10^{-4}$	**0.00890**	*SQLE;SEC24B;GGPS1;NFYA;TGS1;CYP51A1;HMGCR;SEC24D;KPNB1;FDFT1*

M02	Chromatin organization	35/208	$3.76*10^{-15}$	**$1.41*10^{-12}$**	*PHF2;KDM5C;SMARCB1;TRRAP;EHMT2;EHMT1;CHD4;ACTB;PHF21A;NSD1;SAP130;EP400;WDR5;EP300;BRD8;WHSC1;MTA2;KDM6B;BRD1;CREBBP;KDM4B;SMARCC2;KDM2B;SETDB1;SETD1B;USP22;DNMT3A;ARID1A;GATAD2A;HCFC1;SMARCA4;NCOR1;KAT6B;KAT6A;RCOR1*
	Processing of Capped Intron-Containing Pre-mRNA	19/144	$3.55*10^{-7}$	**$8.87*10^{-5}$**	*NUP214;SF3A1;SF3B2;SF3B3;NUP155;FUS;DDX23;SMC1A;PRPF8;SRRM1;NUP93;PRPF6;U2AF2;NUP62;POLR2D;TPR;DHX38;NUP98;SNRNP200*
	Transcription	18/202	$1.02*10^{-4}$	**0.00660**	*GTF3C1;NFIX;POU2F1;EHMT2;CHD4;SSRP1;GATAD2A;SRRM1;POLR3A;POLR1A;U2AF2;POLR2D;TCEB3;UBTF;DHX38;MTA2;TAF4;TAF1*
	PKMTs methylate histone lysines	7/29	$8.03*10^{-5}$	**0.00660**	*SETDB1;EHMT2;NSD1;SETD1B;WDR5;EHMT1;WHSC1*
	Transport of Mature mRNA derived from an Intron-Containing Transcript	9/50	$5.80*10^{-5}$	**0.00660**	*NUP214;NUP93;NUP155;U2AF2;NUP62;TPR;DHX38;NUP98;SRRM1*
	HATs acetylate histones	13/105	$4.91*10^{-5}$	**0.00660**	*BRD1;CREBBP;TRRAP;USP22;ACTB;HCFC1;KAT6B;KAT6A;SAP130;EP400;WDR5;EP300;BRD8*
	Transport of Mature Transcript to Cytoplasm	9/54	$9.84*10-^{5}$	**0.00660**	*NUP214;NUP93;NUP155;U2AF2;NUP62;TPR;DHX38;NUP98;SRRM1*
	mRNA Splicing	13/113	$9.74*10^{-5}$	**0.00660**	*SF3A1;SF3B2;SF3B3;FUS;DDX23;SMC1A;PRPF8;SRRM1;PRPF6;U2AF2;POLR2D;DHX38;SNRNP200*
	Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	13/114	$1.06*10^{-4}$	**0.00660**	*ABCA1;MED1;CREBBP;NCOA6;NRF1;MED26;SREBF2;MED12;MED14;MED24;NCOR1;SIN3A;EP300*
	Transcriptional regulation of white adipocyte differentiation	11/78	$6.57*10^{-5}$	**0.00660**	*MED12;MED1;CREBBP;MED14;MED24;NCOR1;NCOA6;EP300;LPL;MED26;SREBF2*
M16	Axon guidance	11/327	$2.24*10^{-4}$	0.0740	*GSK3B;ARHGEF12;ROCK2;RASA1;KCNQ3;ANK2;ANK3;ARHGEF7;GRIN2B;MYH10;ITGA9*
	Synthesis of PIPs at the early endosome membrane	3/13	$3.84*10^{-4}$	0.0740	*INPP4A;PIKFYVE;PIK3C3*
	CREB phosphorylation through the activation of Ras	3/27	$2.54*10^{-3}$	0.122	*PDPK1;BRAF;GRIN2B*
	Insulin receptor signalling cascade	5/92	0.00191	0.122	*PDPK1;GRB10;PIK3C3;TSC1;MTOR*

EPH-Ephrin signaling	5/94	0.00209	0.122	*ROCK2;RASA1;ARHGEF7;GRIN2B;MYH10*
Sema4D induced cell migration and growth-cone collapse	3/24	0.00187	0.122	*ARHGEF12;ROCK2;MYH10*
Interaction between L1 and Ankyrins	3/29	0.00306	0.131	*KCNQ3;ANK2;ANK3*
Post NMDA receptor activation events	3/35	0.00501	0.143	*PDPK1;BRAF;GRIN2B*
Signaling by Insulin receptor	5/116	0.00497	0.143	*PDPK1;GRB10;PIK3C3;TSC1;MTOR*
PI3K Cascade	4/68	0.00423	0.143	*PDPK1;PIK3C3;TSC1;MTOR*

CREB, cAMP response element binding protein; HATs, histone acetyltransferases; NMDA, N-methyl-D-aspartate; PKMTs, protein lysine methyltransferases; PIPs, phosphatidylinositol phosphates; PI3K, phosphoinositide 3-kinase; snRNP, small nuclear ribonucleo proteins; SREBP, sterol regulatory element-binding proteins; TriC/CCT, TCP1-ring complex or chaperonin containing TCP1. Adjusted P-values with statistical significance are in bold.

Table 3.9 Priority list of 29 essential genes as strong ASD candidates.

Gene	Chr	Start	End	Module	TADA FDR q-value	# High-confidence ASD genes that are co-expressed	Disease associations
BIRC6	2	32357028	32618899	M02	0.47	15	-
CHD1	5	98853985	98928957	M01	0.17	15	*CHD8* has been previously associated with autism
CUL1	7	148697914	148801036	M01	0.49	12	-
DHX29	5	55256245	55307722	M01	0.40	17	-
DVL3	3	184155388	184173610	M02	0.33	10	Robinow syndrome-3, characterized by skeletal abnormalities
EP300	22	41091786	41180077	M02	0.45	13	Rubinstein-Taybi syndrome, characterized by short stature, moderate to severe learning difficulties, distinctive facial features, and broad thumbs and first toes.
EP400	12	131949920	132081102	M02	0.43	9	-
FBXO11	2	47789316	47905793	M01	0.15	17	Associated with chronic otitis media with effusion and recurrent otitis media, a hearing loss disorder, and the ENU knockout of the homologous mouse gene results in the deaf mouse mutant Jeff (Jf)
KDM4B	19	4969113	5153595	M02	0.30	14	-
LDB1	10	102107560	102120453	M02	0.42	14	-
LTN1	21	28928144	28992956	M16	0.37	3	-
MORC3	21	36320189	36386148	M01	0.50	10	-
MYH10	17	8474205	8630761	M16	0.13	3	Essential for normal spine morphology and dynamics. Pharmacologic or genetic inhibition of *Myh10* altered protrusive motility of spines, destabilized their mushroom-head morphology, and impaired excitatory synaptic transmission.
NFIB	9	14081843	14398983	M01	0.45	15	-
PBX1	1	164555584	164899296	M01	0.46	16	-

PHF21A	11	45929323	46121178	M02	0.48	11	-
RFX7	15	56087280	56243266	M01	0.25	17	-
RNF38	9	36336396	36487548	M01	0.41	18	-
SMARCE1	17	40624962	40648508	M01	0.41	12	Meningiomas (brain and spinal cord tumors)
SNW1	14	77717599	77761207	M01	0.44	12	-
STXBP5	6	147204425	147390476	M16	0.37	2	-
SUFU	10	102503987	102633535	M02	0.47	14	Familial Meningioma, Medulloblastoma
TAF4	20	61953469	62065810	M02	0.30	10	Interference of transcription by the binding of TAF4 with expanded polyQ stretches is involved in the pathogenetic mechanisms underlying neurodegeneration.
TANC2	17	63009556	63427699	M02	0.32	14	-
TNPO3	7	128954180	129055173	M01	0.19	17	Mutations found in patients with muscular dystrophy
UTP6	17	31860899	31901765	M01	0.19	12	-
VCP	9	35056064	35073249	M02	0.49	9	Inclusion Body Myopathy with Paget Disease of Bone and Frontotemporal Dementia, Amyotrophic Lateral Sclerosis, Charcot-Marie-Tooth Disease Type 2Y
WHSC1	4	1871424	1982207	M02	0.27	13	Located in the Wolf-Hirschhorn syndrome (WHS) critical region
YTHDC1	4	68310387	68350089	M01	0.48	11	-

Supplementary data

Supplementary data 3.1 List of de novo variants in EGs and NEGs in subjects from the Simons Simplex Collection.

Supplementary data 3.2 List of inherited variants in EGs and NEGs in subjects from the Simons Simplex Collection.

Supplementary data 3.3 Individual mutational burden, essentiality burden score, polygenic risk score and rare deletion burden of subjects from the Simons Simplex Collection.

CHAPTER 4: Essentiality burden score and its application to understanding the genetic architecture of ASD

Introduction

A current model for understanding the genetic etiology of ASD involves the cumulative effects of both common and rare variants including both SNVs and CNVs (de la Torre-Ubieta et al., 2016). A few polygenic methods for investigating the genetic architecture of complex disorders have been proposed (Wray et al., 2014). Polygenic risk score (PRS) aims to provide insight into the genetic architecture captured by common SNVs from GWAS (International Schizophrenia et al., 2009). Specifically, PRS for each individual is calculated as the number of risk alleles weighted by their effect sizes estimated from a discovery GWAS (Wray et al., 2014). Since PRS summarizes the individual-level genetic effects among a group of SNVs that do not individually reach genome-wide significance, it can be used to construct disease risk prediction models (Dudbridge, 2013). PRS analyses have been applied in a few recent ASD studies. For example, The Autism Genome Project Consortium performed a two-stage GWAS in a total of 2,705 ASD familes (including 1,404 families in Stage 1 and 1,301 families in Stage 2) (Anney et al., 2012; Anney et al., 2010). Anney et al. found that PRS based on summary statistics from Stage 1 GWAS is a significant predictor of Stage 2 case-control status (Anney et al., 2012), which validated the collective contribution of common variants to ASD risk. Furthermore, based on ASD GWAS on ~5,000 cases and ~5,000 controls performed by Psychiatric Genomics Consortium (Smoller et al., 2013), Weiner et al. compared the

PRS between ~6,400 ASD children and their parents and confirmed that common polygenic variants contribute to ASD risk in addition to strong acting *de novo* events (Weiner et al., 2016). Recent studies on CNVs implicated their significant role in the genetic architecture of ASD (Bucan et al., 2009; Glessner et al., 2009; Griswold et al., 2012; Itsara et al., 2010; Levy et al., 2011; Malhotra and Sebat, 2012; Marshall et al., 2008; Pinto et al., 2010b; Sanders et al., 2011; Sanders et al., 2015; Sebat et al., 2007; Szatmari et al., 2007). Global CNV load, which is defined as the total number of base pairs covered by CNVs per individual, had been suggested to predispose to autism (Girirajan et al., 2013). While PRS captures the cumulative effect of common variants in both EGs and NEGs, based on an increased burden of inherited, rare and damaging mutations in EGs in ASD probands compared to their unaffected siblings (Ji et al., 2016), I suggested that the burden of rare variants in EGs could be an additional polygenic predictor of individual ASD risk.

In the previous chapter, I defined the individual mutation burden as i) the number of alternative alleles below a defined threshold of minor allele frequency (MAF<0.01) and ii) above a defined threshold of the Combined Annotation-Dependent Depletion (CADD) score that measures variant-level deleteriousness by integrating diverse genome annotations (CADD phred-scale>10) (Ji et al., 2016). However, the individual mutation burden did not take into account the magnitude of damaging effects for each variant, and thus may not be an optimal predictor of individual ASD risk. To evaluate the optimal predictor of individual's ASD risk and to improve the power to differentiate unaffected individuals from ASD subjects with different degrees of social and cognitive impairment,

I developed an Essentiality Burden Score (EBS). The EBS is defined as the weighted sum of coding variants in essential genes per individual. The weights are based on measures of variant-level deleteriousness (as a proxy for pathogenicity), minor allele frequency (MAF), and gene-level intolerance scores (Aggarwala and Voight, 2016; Lek et al., 2016; Petrovski et al., 2013) . The optimized weighing scheme for EBS can be found by maximizing the difference in EBS between ASD patients and unaffected siblings in the discovery dataset (1,781 ASD proband-siblings pairs from Simons Simplex Collection, Table 4.1).

Using weights learned in the discovery phase, I calculated EBS for individuals and families in the target dataset (688 trio families in the ASC exome sequencing dataset, Table 4.1) for validation of the association of EBS to ASD using parent-child relationships. To demonstrate how EBS can facilitate a deeper understanding of the genetic architecture of ASD, I compared EBS, PRS and CNV load and evaluate the power of these polygenic methods in predicting ASD disease status. Moreover, I investigated the potential interplay between EBS and rare variants in high-penetrant ASD risk genes such as *NRXN1*.

Results

Optimization of the essentiality burden score (EBS)

I evaluated the performance of EBS under different weighing schemes in the discovery sample (1,781 ASD proband-sibling pairs from SSC). The contributions of CADD score, MAF [non-Finnish European population in ExAc, (Lek et al., 2016)] and intolerance

scores for the performance of EBS were assessed by four metrics: P values and effect sizes from one-sided Wilcoxon signed rank test for higher EBS in ASD probands; percentage of proband-sibling pairs in which probands have higher EBS (accuracy); area under ROC curve (AUC) for the performance of EBS in discriminating between ASD probands and unaffected siblings (Figure 4.2). I observed that both intolerance scores and CADD score contribute to improved performance of EBS, and that MAF contribute minimally to the performance of EBS. The product of the weights from CADD scores and intolerance scores provides the best performance for EBS (Figure 4.2, Table 4.3, Materials and Methods). Under this model, the P value for elevated EBS in 1,781 ASD probands compared to their unaffected siblings is 2.26×10^{-5}, which is a substantial improvement compared with the individual mutational burden described in chapter 3 (P value =0.0021) (Ji et al., 2016). It is notable that although the effect of increased EBS in ASD probands is statistically significant, the predicting power of EBS in differentiating between ASD probands and unaffected siblings is modest (accuracy=0.542, AUC=0.520).

Regression analysis of the effect of EBS on quantitative traits of ASD probands

Next I investigated the relationship between EBS and quantitative traits available for ~2,500 ASD probands in SSC (Supplementary data 3.3). I confirmed the significant effect of EBS on Social Responsiveness Scale (SRS) in male probands (P value=2.32×10^{-6}; Poisson regression) but not in female probands (P value =0.085; Poisson regression), as well as the effect of EBS on verbal and non-verbal IQ (P value $<2 \times 10^{-16}$ for both verbal and non-verbal IQ; Poisson regression). Moreover, increased EBS in

probands is associated with parents' age at their births, with a stronger effect on the age of fathers (P value =0.00414; Poisson regression) than mothers (P value =0.0109; Poisson regression). When EBS was dissected by the contributions of *de novo* or inherited mutations separately, both *de novo* and inherited mutations contribute to this effect on the age of fathers (P value =1.11×10^{-8} for *de novo*, P value = 0.0384 for inherited; Poisson regression) and mothers (P value = 0.00849 for *de novo*, P value = 0.0283 for inherited; Poisson regression). EBS in the probands is not associated with probands' head circumference (P value = 0.695; Poisson regression) and number of miscarriages in the family (P value = 0.40; Poisson regression).

The extension of polygenic transmission disequilibrium test to EBS

It is expected that a child inherits half of the variants from each of their parents, thus the child's expected EBS is the average EBS of their parents. However, in a collection of trio families in which the children are affected with a trait or disease (such as ASD), the association between EBS and the trait introduces deviation of children's EBS from the expected value. To further validate the association of EBS to ASD using parent-child relationships in the ASC exome sequencing dataset, I extended the polygenic transmission disequilibrium test (Weiner et al., 2016) to EBS. The extension of the transmission disequilibrium test to EBS is equivalent to testing the difference between affected children's EBS and the means of paternal and maternal EBS. Using exome sequencing data of 688 ASD trio families from Autism Sequencing Collaboration, I compared the EBS of ASD children with their unaffected parents (Figure 4.3, Supplementary data 4.1). I found that the EBS of ASD children were greater than the

paternal EBS (P value =0.0136; paired two-sample t-test, one-sided), but not the maternal EBS (P value = 0.6632; paired two-sample t-test, one-sided). In addition, the maternal EBS were marginally greater than paternal EBS in these families (P value = 0.0524; paired two-sample t-test, one-sided). I did not detect significantly increased EBS of ASD children compared to both parents (P value = 0.3675; one-sided two-sample t-test). I found that the ASD children's EBS is marginally increased compared to the means of paternal and maternal EBS (P value = 0.054, paired two-sample t-test, one-sided), further supporting the association between increased burden of deleterious variants in EGs and ASD.

The independent contributions of EBS, polygenic risk score (PRS) and rare deletion burden (RDB) for ASD risk prediction

The cumulative effect of rare variants in EGs is not the only factor that determines ASD disease status, as polygenic effect of common variants, burden of copy number variations (CNVs) and rare variants with high penetrance can be additional genetic factors that predict ASD risk. To test this hypothesis, I took advantage of available exome sequencing and SNP genotyping data of 701 SSC families for which both EBS and PRS could be estimated. The EBS of the 701 SSC proband-sibling pairs were calculated using called variants in their exomes. Using the SNP genotyping data of 701 SSC proband-sibling pairs, I calculated their PRS based on the summary statistics in a GWAS on 5,305 ASD cases and 5,305 psuedocontrols performed by the Psychiatry Genomics Consortium (Smoller et al., 2013). A low negative correlation was observed between the EBS and PRS of 701 SSC ASD probands (Pearson correlation = -0.318, Figure 4.4), suggesting

that the effects of EBS and PRS were independent to each other. Based on a list of rare CNVs in ~2,500 SSC ASD families (Table 4.2) (Sanders et al., 2015), I calculated the rare deletion burden (RDB) of these individuals. RDB was not correlated with either EBS (Pearson correlation=0.042) or PRS (Pearson correlation=-0.056). To evaluate and compare the performance of EBS, PRS and RDB in predicting ASD affected status, I performed logistic regression using the EBS, PRS and RDB of 670 proband-sibling pairs as independent variables. The performance of regression models based combinations of the three metrics considered (EBS, PRS, and RDB) was evaluated by area under ROC curve (AUC) (Figure 4.5). The performance of EBS and that of RDB alone was modest (EBS: AUC=0.514; RDB: AUC=0.515). The performance of PRS is inflated (PRS: AUC=0.806), probably because the ASD GWAS from PGC included overlapping subjects from the SSC cohorts. Regardless of the inflated performance of PRS, EBS and RDB still improved the prediction power when incorporated together with PRS (EBS+PRS: AUC=0.831; RDB+PRS: AUC=0.811). The best performance was achieved when the regression model included all of the three metrics (EBS+RDB+PRS: AUC=0.833) (Figure 4.5, Table 4.4). The regression coefficients of PRS, EBS and RDB were all significant (PRS: P value $<2 \times 10^{-16}$; EBS: P value $=1.28 \times 10^{-12}$; RDB: P value =0.0131). The effect of EBS on ASD affected status was stronger than that of RDB (EBS: coefficient=0.518; RDB: coefficient=0.333). Overall, the results suggest that EBS, PRS and RDB capture different aspects of the genomic landscape of ASD and provide more power to predict ASD risk when combined and considered together.

The interplay between EBS and *NRXN1* variants with a major effect

Next I hypothesized that high-penetrant alleles in ASD risk genes may play an important role in some ASD individuals with lower EBS. *NRXN1* is a synaptic cell-adhesion molecule and a component of the trans-synaptic complex, which is crucial for normal synapse formation and function. *NRXN1* is also an unpublished EG (personal communication with Mark V. Fuccillo) that is significantly associated with ASD [TADA FDR=2.31 $\times 10^{-7}$, (Sanders et al., 2015)]. To investigate the potential interplay between rare variants in *NRXN1* and EBS, I compiled a list of SSC families with rare CNVs and SNVs in *NRXN1* (Figure 4.6). Firstly, I extracted 7 ASD families from SSC with rare exonic deletions in *NRXN1* in at least one child, among which 6 are quartet families (Table 4.5). All of these 6 rare exonic deletions in *NRXN1* were in probands (4 inherited and 2 *de novo*) and none was present in siblings only, indicating that these deletions could have strong effects and lead to ASD. Among the 7 probands with rare exonic deletions in *NRXN1*, 5 had lower EBS compared to their unaffected siblings. Secondly, I extracted 19 ASD quartet families from 1,781 SSC families in which rare functional mutations in *NRXN1* were present in at least one child (Figure 4.6, Table 4.6). There were 14 unique heterozygous mutations in *NRXN1* in these 19 families (13 inherited and 1 *de novo*). With the assumption that *NRXN1* mutations present in probands have stronger penetrance than those mutations only observed in unaffected siblings, I found that ASD probands with high-penetrance *NRXN1* mutations are more likely to have lower EBS compared to their siblings (odds ratio=11.25, P value=0.0495; one-sided Fisher's exact test) (Table 4.7). These findings suggest that in some ASD patients, alleles in high-penetrance ASD risk genes such as *NRXN1* play an important role in the development of ASD, regardless of their lower mutational burden in EGs.

Discussion

I developed the essentiality burden score (EBS) by giving weights to different variants according to their functional impact, which provides greater power to separate ASD probands from their unaffected siblings. I validated the effectiveness of EBS in an independent cohort (ASC) by extending the polygenic transmission disequilibrium test to EBS and showed that the EBS of ASD probands tends to be higher than the average EBS of their parents. I demonstrated the independent contributions of EBS, polygenic risk score (PRS) and rare deletion burden (RDB) for ASD risk prediction. Finally, I observed a potential interplay between EBS and *NRXN1* variants with a major effect in ASD probands from Simons Simplex Collection.

Since the era of genome-wide association studies (GWAS), the large proportion of the heritability of complex disorders such as ASD unexplained by GWAS, i.e. the "missing heritability" (Manolio et al., 2009; McCarthy et al., 2008), remains as a key question to be answered. With the advancement of next generation sequencing technology, many recent ASD studies have been successful in identifying many more ASD candidate genes by focusing on rare and *de novo* variants with a large functional impact (Iossifov et al., 2014; Iossifov et al., 2012; Levy et al., 2011; Sanders et al., 2011; Sanders et al., 2012). Besides common variants in GWAS or rare variants with large genetic effect, my study focused on an important but less explored territory: low-frequency or rare variants with small to intermediate effect in EGs (Figure 4.7). The EBS captures the cumulative effect of a large number of rare deleterious variants in EGs. I found significantly elevated EBS in ASD probands and an independent contribution of EBS to ASD risk prediction, which

suggests that EBS analysis, along with PRS analysis that captures cumulative effect of common variants, is useful in identifying the "still-missing" heritability of ASD and thus important for a full understanding of the genetic landscape of ASD. Moreover, my results on elevated EBS in unaffected mothers compared to unaffected fathers in ASD trio families from the ASC cohort further support the "female protective model" in ASD, for which several studies have gathered reinforcing genetic evidence showing a higher burden of *de novo* loss-of function mutations (Iossifov et al., 2012) and CNVs (Jacquemont et al., 2014; Levy et al., 2011; Sanders et al., 2011; Sanders et al., 2015) in female ASD probands.

It is not surprising that the predictive power of EBS alone on ASD risk is only moderate. Firstly, due to the complexity of the genomic architecture of ASD (de la Torre-Ubieta et al., 2016), we may expect that it is not a single genetic predictor but the joint contribution of a broad spectrum of genetic variants (Figure 4.7) that determines the genetic risk of ASD. This notion is supported by my results that i) the performance of ASD predicting models including all three genetic predictors (i.e. EBS, PRS and RDB) is superior to the models with only a single predictor and ii) high-penetrant *NRXN1* mutations are associated with lower EBS in ASD probands. Secondly, for complex genetic disorders we may expect an influence of environmental factors, and therefore the predictive value of a single genetic risk factor is limited (Wray et al., 2014; Wray et al., 2010). Other identified non-genetic factors in ASD such as parental age, prenatal stress or infection, maternal Zn^{2+}-deficiency and maternal exposure of toxins (Grabrucker, 2012), should also be taken into consideration in order to increase the predictive ability of genetic predictors

(Wray et al., 2014). Thirdly, there are still limitations in the current model for EBS, which has not taken into consideration variants in the X chromosome (that harbors 89 EGs) and non-coding variants involved in regulating the expression of EGs. Finally, it is expected that EBS will achieve a better predictive value for ASD risk as the catalog of EGs continues to grow.

Materials and Methods

ASC exome sequencing data

Exome sequence reads of 3,417 individuals in ARRA Autism Sequencing Collaboration (ASC) were downloaded from dbGaP (https://www.ncbi.nlm.nih.gov/gap; dbGaP Study Accession: phs000298.v1.p1). The reads were aligned to GRCh37/hg19 human genome with Burrows-Wheeler Aligner (BWA) 0.7.10 (Li and Durbin, 2009). Variant calling was performed with the Genome Analysis Toolkit (GATK) (3.2.2) with default protocols (McKenna et al., 2010). Variant filters include i) passed the GATK variant quality score recalibration (VQSR) filter, ii) call rate>=90% and c) Hardy Weinberg equilibrium P value $>1\times 10^{-6}$. Genotype filters include i) read depth (DP) >=8 and ii) genotype quality (GQ) >=20. The final variant call set included 934,511 variants with an average call rate of 99.2% and a transition / transversion ratio of 2.75. From the 3,417 individuals, I extracted 688 trio families with two unaffected parents and a child affected with ASD.

SSC SNP genotyping data

I obtained Illumina 1M SNP genotyping data of 4,753 individuals (1,223 probands and 3,530 unaffected family members) in SSC. Invariant markers, duplicated variants, non-

autosomal variants and variants with call rate <90% and Hardy Weinberg equilibrium P value < 1×10^{-7} were removed. The final variant call set included 886,001 variants with an average call rate of 99.96%. The chromosomal positions of the variants were lifted over from hg18 to hg19 using the LiftOver tool from UCSC genome browser (https://genome.ucsc.edu). SNP array data were available for 701 SSC proband-sibling pairs among the 1,781 quartet families with exome sequencing data.

Definition and optimization of the essentiality burden score (EBS)

To calculate the EBS, suppose there are n functional variants (loss-of-function or missense) in a whole exome sequence or whole genome sequence dataset, I_{ij} indicates whether variant *j* exists in individual *i*. The EBS for individual *i* is defined as the weighted sum of the number of risk alleles in an individual:

$$EBS_i = \sum_{j=1}^{n} (W_j I_{ij})$$

For each functional variant, the weighing metric (*W*) combines evidence from i) minor allele frequency (MAF) in ExAC (European ancestry) (Lek et al., 2016), ii) variant-level deleteriousness quantified by CADD score (Kircher et al., 2014) and iii) gene-level intolerance scores including Residual Variation Intolerance Score (RVIS) (Petrovski et al., 2013), the probability of being loss-of-function intolerant (pLI) (Dickinson et al., 2016)and genic intolerance based on sequence context (Aggarwala's score) (Aggarwala and Voight, 2016) .

$$W_{CADD} = CADD\ phred\text{-}scale\ score$$

$$W_{MAF} = [Beta\ (MAF;\ 1,25)]^2$$ (Wu et al., 2011)

$$W_{intolerance} = 100-(RVIS\ percentile+pLI\ percentile+Aggarwala\ percentile)/3$$

$$W= (W_{CADD}*a_1 + W_{MAF}*a_2)*W_{intolerance}\text{^}a3$$

The objective of optimization was to learn the weights (a_1, a_2 and a_3), using the discovery dataset, that maximally differentiates ASD cases from family controls. To achieve this, I used annotated variants in the discovery sample [1,781 ASD quartet families from SSC, specifically, variants reported in (Iossifov et al., 2014) and (Krumm et al., 2015)], to identify a combination of weighing metrics by maximizing the performance of EBS in differentiating between ASD probands and unaffected siblings. The performance of EBS were assessed by four metrics: P values and effect sizes from one-sided Wilcoxon signed rank test for higher EBS in ASD probands; percentage of proband-sibling pairs in which probands have higher EBS (accuracy); area under ROC curve (AUC) for the performance of EBS in discriminating between ASD probands and unaffected siblings.

Calculation of ASD polygenic risk score (PRS)

I downloaded GWAS results of a meta-analysis of 5,305 ASD-diagnosed cases and 5,305 psuedocontrols constructed from untransmitted parental chromosome performed by the Psychiatric Genomics Consortium (Robinson et al., 2016). Using PRSice 1.25 (Euesden et al., 2015), I calculated the PRS of 701 SSC proband-sibling pairs at a P value threshold of 0.5, at which 1,223 probands were maximally differentiated from 3,530 unaffected family members (Figure 4.1).

Calculation of rare deletion burden (RDB)

I acquired a list of rare CNVs (with population frequency <= 0.1%) predicted from Illumina Omni2.5 SNP genotyping data of 2,591 SSC ASD families (Table 4.2) (Sanders et al., 2015). The rare deletion burden (RDB) was defined as the total number of base pairs covered by rare deletions across an individual genome (Girirajan et al., 2013).

Figures

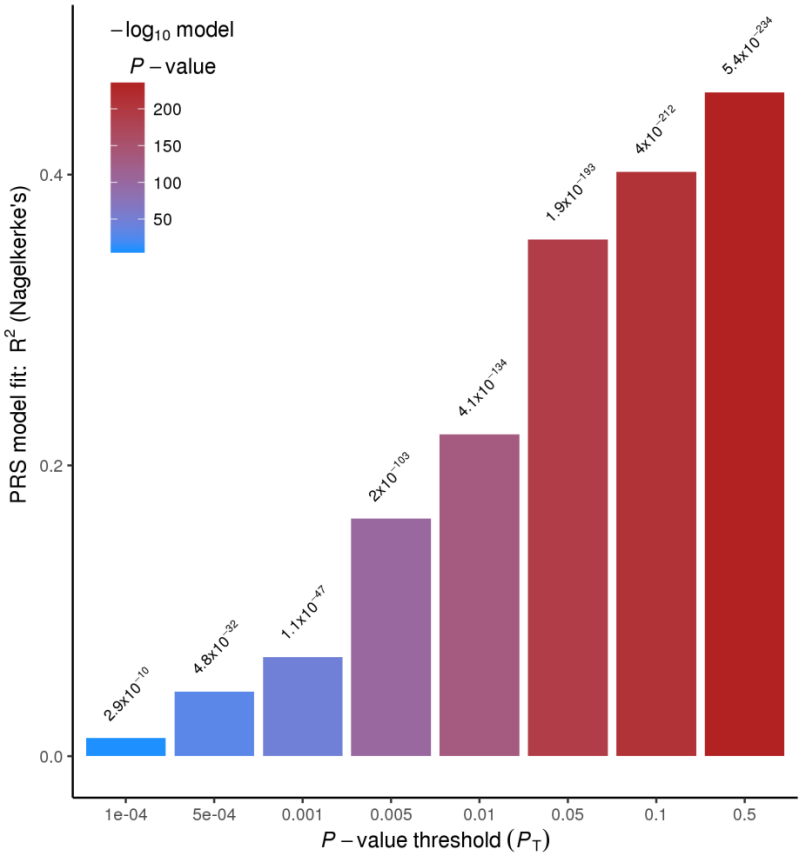

Figure 4.1 PRS model fit across multiple GWAS P-value thresholds. The figure was generated by PRSice 1.25. For each GWAS p-value threshold (x-axis), a regression was performed to test the association between the PRS and ASD affected status of 1,223 SSC probands and 3,530 unaffected family members. The R^2 and p-value for each regression model is shown.

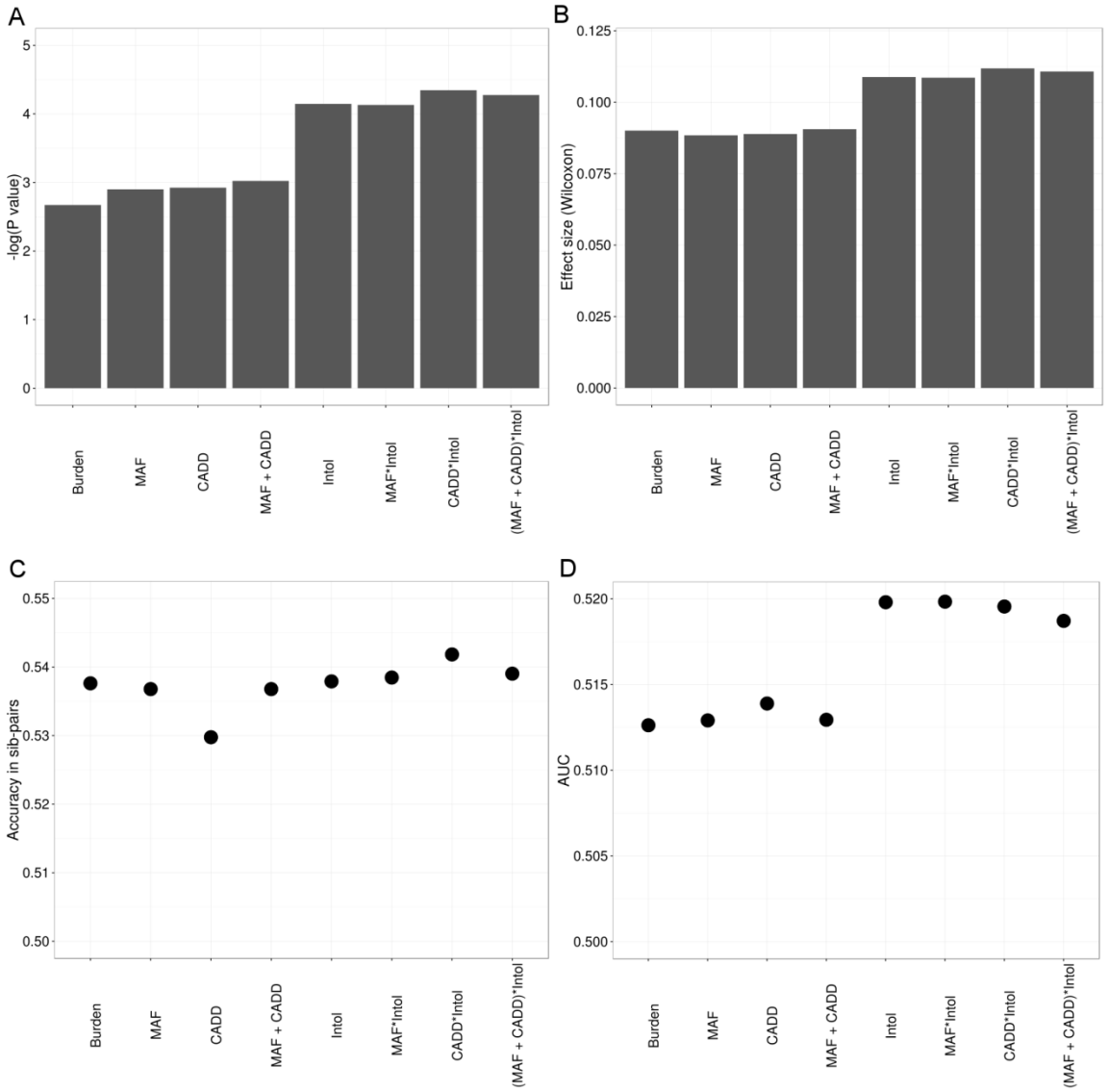

Figure 4.2 Optimization of parameters for EBS. Performance of EBS in the discovery sample (1,781 ASD proband-sibling pairs from SSC) was evaluated with four metrics: **(A)** p-values **(B)** effect sizes from one-sided Wilcoxon signed rank test for higher EBS in ASD probands, **(C)** percentage of proband-sibling pairs in which probands have higher EBS, **(D)** area under ROC curve for the performance of EBS in discriminating between ASD probands and unaffected siblings.

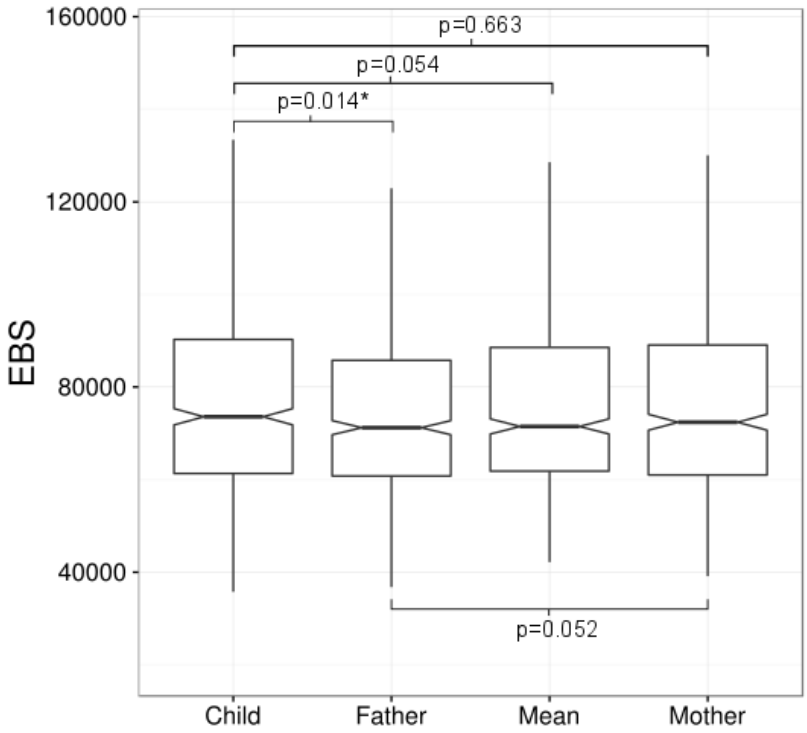

Figure 4.3 Essentiality burden scores in ASD trio families. The boxplots indicate the distribution of essentiality burden score (EBS) of ASD children, mothers and fathers in 685 trio families from ARRA Autism Sequencing Collaboration. “Mean” stands for the mean of paternal and maternal EBS. One-sided paired two-sample t-tests for increased EBS in ASD children (top three p-values) and for increased EBS in mothers (bottom p-value) were performed. *p-value<0.05.

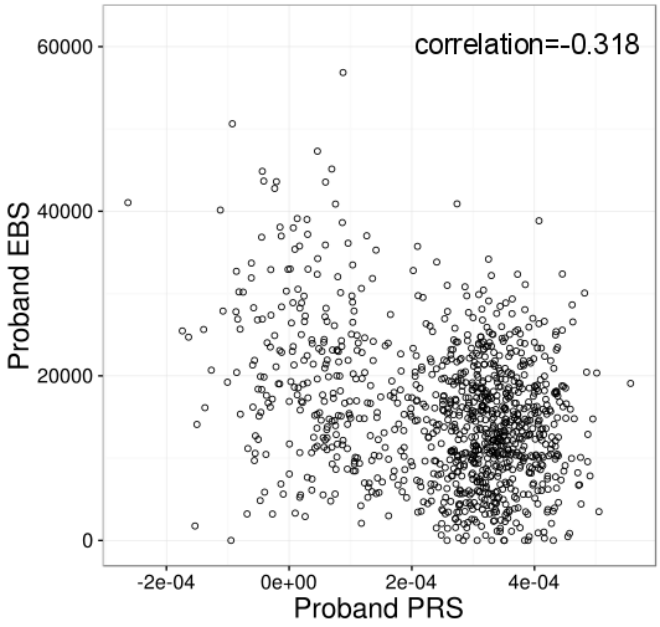

Figure 4.4 The EBS and PRS of 701 ASD probands. The Pearson correlation between EBS and PRS is shown.

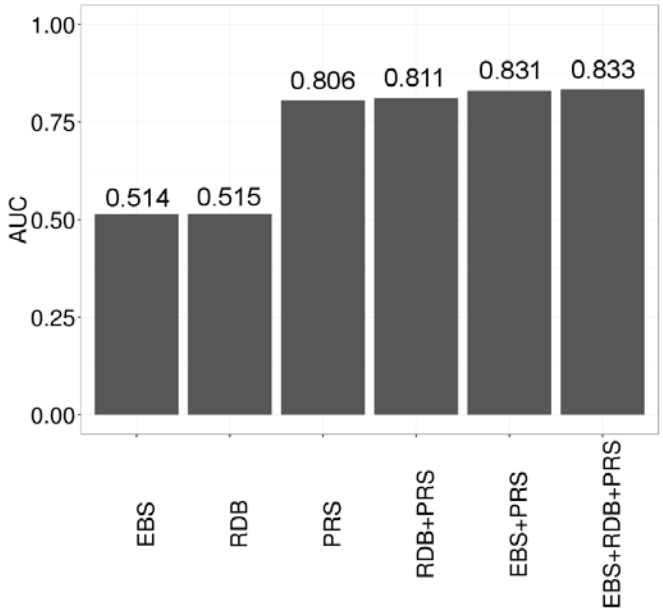

Figure 4.5 Performance of ASD prediction models. Multivariate logistic regression was performed to predict the affected status of 670 proband-sibling pairs. Combinations of three individual-level metrics [i.e. essentiality burden score (EBS), rare deletion burden (RDB), and polygenic risk score (PRS)] were used as independent variables in the regression models. The area under ROC curve for each of the six regression models is shown.

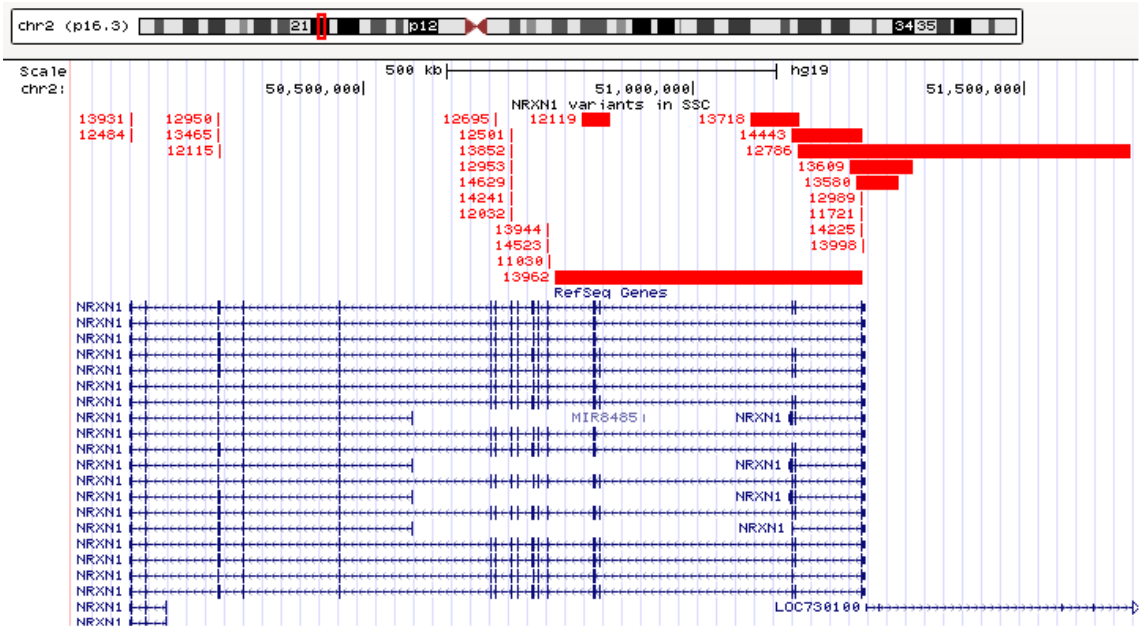

Figure 4.6 Rare exonic deletions and SNVs in *NRXN1* in SSC. UCSC Genome Browser (http://genome.ucsc.edu/) view of 7 families with rare exonic deletions (long red bars) and 19 families with rare exonic SNVs (short red bars) in *NRXN1* with SSC family ID is shown.

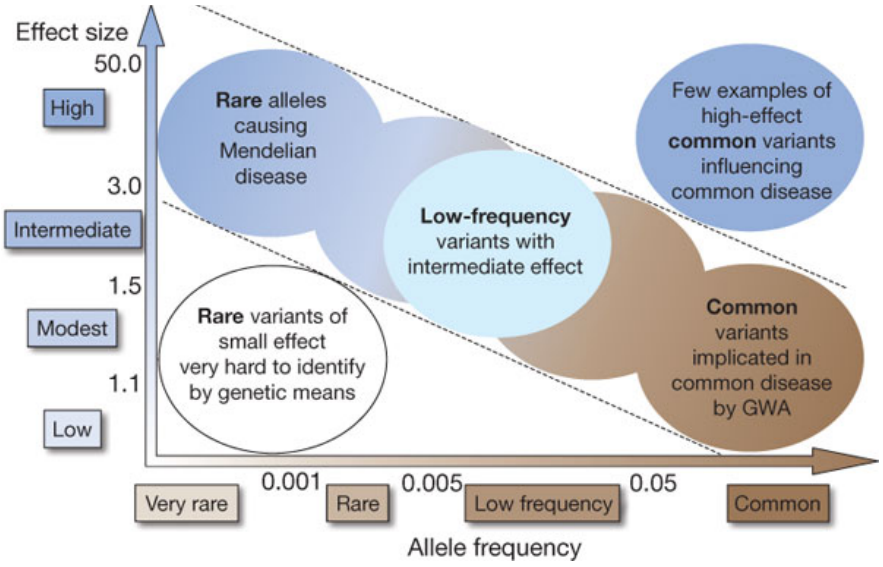

Figure 4.7 Spectrum of complex disease risk variants by allele frequency and effect size. This figure is from ref (Manolio et al., 2009), which is adapted from ref (McCarthy et al., 2008).

Tables

Table 4.1 Datasets involved in Chapter 4.

Cohort	Type	Platform	Sample size	Source
SSC	Rare variants (from whole exome sequencing)	Illumina HiSeq 2000	2,368 trio and quartet families	Iossifov et al. 2014; Krumm et al. 2015
SSC	SNP genotyping	Illumina 1M	701 quartet families	SSC
SSC	Rare CNVs	Illumina 1M and Omni2.5	2,591 trio and quartet families	Sanders et al. 2015
ASC	Whole exome sequencing	Illumina HiSeq 2000	688 trio families	ASC (dbGaP)

The intersection between 2,369 families with rare variants and 2,591 families with rare CNVs are 2,181. 701 families with SNP genotyping data are a subset of 2,368 families with rare variants.

Table 4.2 Statistics of rare CNVs in 2,591 SSC ASD families.

	Deletion			Duplication		
	Exonic/Splicing	Intronic/UTR	Intergenic	Exonic/Splicing	Intronic/UTR	Intergenic
Both	173	15	184	504	0	65
EG	1379	1874	1271	1031	400	336
NEG	2144	2788	1611	1496	579	420
Unknown	8231	6518	10866	4747	1390	3226
Total	11927	11195	13932	7778	2369	4047

Table 4.3 The performances of different models of essentiality burden score.

Model	P	Effect Size	Accuracy	AUC	a1	a2	a3
Burden	0.0021	0.0901	0.5376	0.5126	.	.	.
MAF	0.0013	0.0884	0.5368	0.5129	0	1	0
CADD	0.0012	0.0889	0.5298	0.5139	10	0	0
MAF+CADD	0.0009	0.0906	0.5368	0.5129	10	1	0
Intol	7.15E-05	0.1088	0.5379	0.5198	.	.	.
MAF*Intol	7.41E-05	0.1086	0.5385	0.5198	0	1	1
CADD*Intol	4.52E-05	0.1118	0.5418	0.5196	10	0	1
(MAF+CADD)*Intol	5.31E-05	0.1108	0.5390	0.5187	10	1	1

Table 4.4 Regression analysis to predict ASD affected status.

	Estimate	Std. Error	Z value	p-value
EBS	0.51791	0.07298	7.096	1.28×10^{-12} *
RDB	0.33346	0.13444	2.480	0.0131 *
PRS	1.72517	0.09491	18.176	$< 2 \times 10^{-16}$ *
Intercept	0.08912	0.06945	1.283	0.199

Logistic regression was performed for the essentiality burden score (EBS), rare deletion burden (RDB), and polygenic risk score (PRS) of 670 proband-sibling pairs from SSC. * p-value <0.05.

Table 4.5 ASD families from SSC with rare exonic deletions in *NRXN1*.

Family	Chr	Start	Stop	Inheritance	EBS proband	EBS sibling	EBS difference
12119	2	50831734	50873107	*de novo*	15598.29	NA	NA
12786	2	51158351	51661515	mother-both	24412.65	33687.31	-9274.67
13580	2	51247294	51311532	*de novo*	23413.55	18604.00	4809.55
13609	2	51236179	51332477	father-pro	459.63	1132.83	-673.20
13718	2	51087308	51161424	mother-pro	10865.98	17306.43	-6440.45
13962	2	50790714	51256013	*de novo*	8644.48	13599.03	-4954.55
14443	2	51149414	51255832	father-both	1798.76	1808.08	-9.32

Table 4.6 ASD families from SSC with rare/damaging SNVs or indels in *NRXN1*.

Family	Chr	Start	Stop	Ref	Alt	Inherit-ance	EBS proband	EBS sibling	EBS difference
13465	2	50280493	50280493	G	-	mo-pro	11078.04	19319.46	-8241.42
12115	2	50282085	50282085	-	T	fa-sib	8444.06	7978.35	465.71
12989	2	51253590	51253590	G	A	fa-both	18550.44	14013.83	4536.62
12501	2	50724605	50724605	A	T	*de novo*	14371.08	14798.24	-427.15
12032	2	50724817	50724817	G	A	mo-pro	23199.61	23704.83	-505.22
14241	2	50724817	50724817	G	A	fa-pro	18231.69	15067.18	3164.52
14629	2	50724817	50724817	G	A	fa-sib	15127.46	22051.87	-6924.41
12953	2	50724817	50724817	G	A	fa-pro	13878.68	16416.71	-2538.03
13852	2	50724817	50724817	G	A	mo-sib	729.02	0	729.02
13998	2	51255218	51255218	C	T	fa-both	36937.25	36923.27	13.98
13931	2	50149233	50149233	C	T	mo-pro	13900.33	16532.63	-2632.3
11030	2	50780151	50780151	T	A	mo-sib	13235.1	4629.40	8605.70
14523	2	50779938	50779938	G	T	mo-sib	24330.59	22213.44	2117.15
12484	2	50149314	50149314	G	A	mo-pro	19311.12	16490.36	2820.7
13944	2	50779784	50779784	A	C	mo-sib	27655.17	26322.5	1332.67
12950	2	50280477	50280477	T	A	mo-both	24713.45	27057.88	-2344.44
12695	2	50699532	50699532	G	C	mo-both	19295.37	23723.99	-4428.62
14225	2	51253608	51253608	C	T	fa-both	22945.78	24235.69	-1289.91
11721	2	51253608	51253608	C	T	fa-pro	21189.56	21563.92	-374.36

All listed variants have CADD score>10. mo, mother; fa, father.

Table 4.7 Relationship between mutations in *NRXN1* and EBS in ASD quartet families.

***NRXN1* mutation**	**Lower EBS in probands**	**Higher EBS in probands**	**Odds Ratio**	**p-value**
in proband only	6	2	15	0.0513
in both	3	2	7.5	0.1970
in proband	9	4	11.25	0.0495 *
in sibling only	1	5	NA	NA

One-sided Fisher's exact test was performed with "in sibling only" group as control. *p-value <0.05.

Supplementary data

Supplementary data 4.1 Essentiality burden score of subjects from the Autism Sequencing Collection.

CHAPTER 5: Conclusion and future directions

The goal of this dissertation is to systemically characterize human essential genes (EGs) and investigate the role of EGs in neurodevelopmental disorders such as ASD. From the analysis of the most comprehensive set of human EGs to date, this study demonstrated that i) EGs are not only relevant but also important for both Mendelian and complex diseases and ii) mutational load in EGs plays a significant role in the genetic basis of a neurodevelopmental disorder, ASD. Therefore, the analysis of EGs can serve as an important step for both interpretation and prioritization of ASD risk alleles, as well as a full understanding of the genetic landscapes of ASD and possibly more complex diseases. In light of the findings and discussions in previous chapters, I will discuss possible future directions for the analysis of EGs.

Firstly, the extension of EG analysis to other complex diseases will further validate the key findings in this dissertation and lead to new insights into the role of the "essentialome" in human disease etiology. According to my results in Chapter 2, EGs are especially enriched among genes associated with early onset diseases, which is consistent with EGs' vital role during embryonic development. Neurodevelopmental disorders (e.g. ASD, intellectual disabilities, developmental co-ordination disorder and attention-deficit/hyperactivity disorder) are a group of conditions that manifest early in the development of children. Moreover, some neurodevelopmental disorders may share similar genetic risk factors (Smoller et al., 2013) and they frequently co-occur (Leitner, 2014; Tonnsen et al., 2016). Therefore, neurodevelopmental disorders other than ASD

are plausible targets for future analyses of EGs. However, we cannot rule out the possibility that EGs could also play a role in later onset diseases, thus a comparison of the contributions of EGs in early onset and late onset diseases is warranted.

Secondly, multiple lines of evidence suggest that genetic variants in functional elements within noncoding genomic regions play an important role in complex diseases (Zhang and Lupski, 2015). Since the coding regions of EGs are intolerant to deleterious mutations, it is expected that the regulatory elements of EGs may also have a distinct mutational spectrum. This notion is supported by Lek et al., who observed that the most highly mutational constrained genes (which overlap greatly with EGs) are depleted for expression quantitative loci (eQTL) (Lek et al., 2016). It implies that the regulatory elements of EGs are less redundant and functional variants in these elements are more likely to contribute to disease-related phenotypes. Therefore, a better understanding of the regulatory elements of EGs will improve our ability to interpret the functional consequences of non-coding variants discovered from whole genome sequencing studies. Moreover, by including variants in noncoding regulatory regions of EGs in the individual essentiality burden score (EBS), the power of EBS to predict individual ASD risk will likely be enhanced. In order to achieve this goal, one of the key problems to be solved is to credibly identify the regulatory elements for each EG. Multiple approaches can be applied independently or together to identify the regulatory elements of EGs. For example, analysis of eQTLs establishes the association between non-coding variants and gene expression levels by combining whole genome sequencing data and RNA sequencing data of the same individuals (Albert and Kruglyak, 2015). Chromosome

conformation capture followed by massively parallel sequencing (Hi-C) allows whole genome mapping of long range physical interactions (Lieberman-Aiden et al., 2009). With careful data quality control and selection of predictive models, eQTL and/or Hi-C data can be used to systematically assign regulatory elements to EGs.

Last but not least, compiling a complete set of ~6,000 putative human EGs will be an achievable goal in the near future, as the genetic community including the International Mouse Phenotyping Consortium (IMPC) continues to produce and phenotype knockout mouse lines for the remainder of the ~20,000 genes in the mouse genome. Before this task is completed, an alternative approach to expand the current catalog of human EGs is to take advantage of essentiality in model organisms other than mouse. Large scale phenotypic analysis of mutant strains has been performed in *S. cerevisiae* (Giaever et al., 2002; Winzeler et al., 1999), *C. elegans* (Clark et al., 1988; Johnsen and Baillie, 1991), *D. melanogaster* (Boutros et al., 2004; Dietzl et al., 2007; Kamath et al., 2003) and *D. rerio* (zebrafish) (Amsterdam et al., 2004). Human orthologs of EGs in these organisms are also plausible candidates for human EGs. However, since these organisms are not as close evolutionary relatives to human as mouse, EGs in these organisms may not be essential in mammals because of commonly occurring gene duplication during evolution (Holland et al., 1994). Therefore, extra care should be taken when inferring human EGs from these organisms. For example, human orthologs of EGs in multiple organisms could be highly conserved during evolution and thus are more likely to be EGs in human. Overall, a larger or eventually complete catalog of human EGs will further deepen our

understanding of basic biological processes and their contribution of EGs to human disease.

BIBLIOGRAPHY

Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Molecular autism *4*, 36.

Aggarwala, V., and Voight, B.F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nature genetics *48*, 349-355.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature reviews Genetics *16*, 197-212.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders: DSM-5.

Amsterdam, A., Nissen, R.M., Sun, Z., Swindell, E.C., Farrington, S., and Hopkins, N. (2004). Identification of 315 genes essential for early zebrafish development. Proceedings of the National Academy of Sciences of the United States of America *101*, 12792-12797.

Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., Bolshakova, N., Bolte, S., Bolton, P.F., Bourgeron, T., *et al.* (2012). Individual common variants exert weak effects on the risk for autism spectrum disorderspi. Human molecular genetics *21*, 4781-4792.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Sykes, N., Pagnamenta, A.T., *et al.* (2010). A genome-wide scan for common alleles affecting risk for autism. Human molecular genetics *19*, 4072-4082.

Bailey, A., Lecouteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., and Rutter, M. (1995). Autism as a Strongly Genetic Disorder - Evidence from a British Twin Study. Psychol Med *25*, 63-77.

Belinson, H., Nakatani, J., Babineau, B.A., Birnbaum, R.Y., Ellegood, J., Bershteyn, M., McEvilly, R.J., Long, J.M., Willert, K., Klein, O.D., *et al.* (2016). Prenatal beta-catenin/Brn2/Tbr2 transcriptional cascade regulates adult social and stereotypic behaviors. Molecular psychiatry.

Bellen, H.J., Levis, R.W., Liao, G.C., He, Y.C., Carlson, J.W., Tsang, G., Evans-Holm, M., Hiesinger, P.R., Schulze, K.L., Rubin, G.M., *et al.* (2004). The BDGP gene disruption project: Single transposon insertions associated with 40% of Drosophila genes. Genetics *167*, 761-781.

Bergmiller, T., Ackermann, M., and Silander, O.K. (2012). Patterns of Evolutionary Conservation of Essential Genes Correlate with Their Compensability. PLoS genetics *8*.

Blomen, V.A., Majek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., *et al.* (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092-1096.

Boutros, M., and Ahringer, J. (2008). The art and design of genetic screens: RNA interference. Nature reviews Genetics *9*, 554-566.

Boutros, M., Kiger, A.A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S.A., Paro, R., Perrimon, N., and Heidelberg Fly Array, C. (2004). Genome-wide RNAi analysis of growth and viability in Drosophila cells. Science *303*, 832-835.

BrainSpan: Atlas of the Developing Human Brain [Internet]. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. ©2011. Available from: http://brainspan.org

Bucan, M., Abrahams, B.S., Wang, K., Glessner, J.T., Herman, E.I., Sonnenblick, L.I., Alvarez Retuerto, A.I., Imielinski, M., Hadley, D., Bradfield, J.P., *et al.* (2009). Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. PLoS genetics *5*, e1000536.

Buxbaum, J.D., Daly, M.J., Devlin, B., Lehner, T., Roeder, K., and State, M.W. (2012). The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron *76*, 1052-1056.

Cantor, R.M., Kono, N., Duvall, J.A., Alvarez-Retuerto, A., Stone, J.L., Alarcon, M., Nelson, S.F., and Geschwind, D.H. (2005). Replication of autism linkage: fine-mapping peak at 17q21. American journal of human genetics *76*, 1050-1056.

Chakravarti, A., and Turner, T.N. (2016). Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. BioEssays : news and reviews in molecular, cellular and developmental biology *38*, 578-586.

Chang, J., Gilman, S.R., Chiang, A.H., Sanders, S.J., and Vitkup, D. (2015). Genotype to phenotype relationships in autism spectrum disorders. Nature neuroscience *18*, 191-198.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q.N., Wang, Z.C., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics *14*.

Choi, J., Shooshtari, P., Samocha, K.E., Daly, M.J., and Cotsapas, C. (2016). Network Analysis of Genome-Wide Selective Constraint Reveals a Gene Network Active in Early Fetal Brain Intolerant of Mutation. PLoS genetics *12*, e1006121.

Christensen, D.L. (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012 (vol 65, pg 1, 2016). Mmwr-Morbid Mortal W *65*, 404-404.

Clark, D.V., Rogalski, T.M., Donati, L.M., and Baillie, D.L. (1988). The Unc-22(Iv) Region of Caenorhabditis-Elegans - Genetic-Analysis of Lethal Mutations. Genetics *119*, 345-353.

Cnv, Schizophrenia Working Groups of the Psychiatric Genomics, C., and Psychosis Endophenotypes International, C. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nature genetics *49*, 27-35.

Constantino, J., and Gruber, C. (2005). The social responsiveness scale manual. Los Angeles: Western. Psychological Services.

Constantino, J.N., Davis, S.A., Todd, R.D., Schindler, M.K., Gross, M.M., Brophy, S.L., Metzger, L.M., Shoushtari, C.S., Splinter, R., and Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. J Autism Dev Disord *33*, 427-433.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., *et al.* (2010). The genetic landscape of a cell. Science *327*, 425-431.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G.M., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., *et al.* (2014). The Reactome pathway knowledgebase. Nucleic acids research *42*, D472-D477.

Dang, V.T., Kassahn, K.S., Marcos, A.E., and Ragan, M.A. (2008). Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. European Journal of Human Genetics *16*, 1350-1357.

de la Torre-Ubieta, L., Won, H.J., Stein, J.L., and Geschwind, D.H. (2016). Advancing the understanding of autism disease mechanisms through genetics. Nature medicine *22*, 345-361.

De Rubeis, S., and Buxbaum, J.D. (2015). Recent advances in the genetics of autism spectrum disorder. Current neurology and neuroscience reports *15*, 36.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., *et al.* (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209-U119.

Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C., and Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics *169*, 1915-1925.

Dickerson, J.E., Zhu, A., Robertson, D.L., and Hentges, K.E. (2011). Defining the role of essential genes in human disease. PloS one *6*, e27368.

Dickinson, M.E., Flenniken, A.M., Ji, X., Teboul, L., Wong, M.D., White, J.K., Meehan, T.F., Weninger, W.J., Westerberg, H., Adissu, H., *et al.* (2016). High-throughput discovery of novel developmental phenotypes. Nature *537*, 508-514.

Dietzl, G., Chen, D., Schnorrer, F., Su, K.C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S., *et al.* (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature *448*, 151-U151.

Doan, R.N., Bae, B.I., Cubelos, B., Chang, C., Hossain, A.A., Al-Saad, S., Mukaddes, N.M., Oner, O., Al-Saffar, M., Balkhy, S., *et al.* (2016). Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell *167*, 341-354 e312.

Domazet-Loso, T., and Tautz, D. (2008). An Ancient Evolutionary Origin of Genes Associated with Human Genetic Diseases. Molecular biology and evolution *25*, 2699-2707.

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. PLoS genetics *9*.

Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends in Genetics *29*, 569-574.

Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A., and Grp, M.G.D. (2005). The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology. Nucleic acids research *33*, D471-D475.

Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: Polygenic Risk Score software. Bioinformatics *31*, 1466-1468.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., *et al.* (2016). The Reactome pathway Knowledgebase. Nucleic acids research *44*, D481-D487.

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proceedings of the National Academy of Sciences of the United States of America *105*, 4323-4328.

Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron *68*, 192-195.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., *et al.* (2014). Ensembl 2014. Nucleic acids research *42*, D749-D755.

Folstein, S., and Rutter, M. (1977). Genetic Influences and Infantile-Autism. Nature *265*, 726-728.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv *http://arxiv.org/abs/1207.3907*.

Gauglerl, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M.L., Manaa, D., Pawitan, Y.D., Reichert, J., *et al.* (2014). Most genetic risk for autism resides with common variation. Nat Genet *46*, 881-885.

Georgi, B., Voight, B.F., and Bucan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS genetics *9*, e1003484.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., *et al.* (2003). Experimental

determination and system level analysis of essential genes in Escherichia coli MG1655. Journal of bacteriology *185*, 5673-5684.

Geschwind, D.H., and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. Current opinion in neurobiology *17*, 103-111.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., *et al.* (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387-391.

Gibson, G. (2012). Rare and common variants: twenty arguments. Nature reviews Genetics *13*, 135-145.

Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron *70*, 898-907.

Girirajan, S., and Eichler, E.E. (2010). Phenotypic variability and genetic susceptibility to genomic disorders. Human molecular genetics *19*, R176-187.

Girirajan, S., Johnson, R.L., Tassone, F., Balciuniene, J., Katiyar, N., Fox, K., Baker, C., Srikanth, A., Yeoh, K.H., Khoo, S.J., *et al.* (2013). Global increases in both common and rare copy number load associated with autism. Human molecular genetics *22*, 2870-2880.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., *et al.* (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature *459*, 569-573.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proceedings of the National Academy of Sciences of the United States of America *104*, 8685-8690.

Grabrucker, A.M. (2012). Environmental factors in autism. Frontiers in psychiatry *3*, 118.

Gratten, J., Visscher, P.M., Mowry, B.J., and Wray, N.R. (2013). Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. Nature genetics *45*, 234-238.

Griswold, A.J., Ma, D., Cukier, H.N., Nations, L.D., Schmidt, M.A., Chung, R.H., Jaworski, J.M., Salyakina, D., Konidari, I., Whitehead, P.L., *et al.* (2012). Evaluation of

copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. Human molecular genetics *21*, 3513-3523.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K., *et al.* (2011). Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. Archives of general psychiatry *68*, 1095-1102.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research *33*, D514-D517.

Harborth, J., Elbashir, S.M., Bechert, K., Tuschl, T., and Weber, K. (2001). Identification of essential genes in cultured mammalian cells using small interfering RNAs. Journal of cell science *114*, 4557-4565.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., *et al.* (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*.

Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., Elison, J.T., Swanson, M.R., Zhu, H., Botteron, K.N., *et al.* (2017). Early brain development in infants at high risk for autism spectrum disorder. Nature *542*, 348-351.

He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D., *et al.* (2013). Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. PLoS genetics *9*.

Herrmann, T., van der Hoeven, F., Grone, H.J., Stewart, A.F., Langbein, L., Kaiser, I., Liebisch, G., Gosch, I., Buchkremer, F., Drobnik, W., *et al.* (2003). Mice with targeted disruption of the fatty acid transport protein 4 (Fatp 4, SlC27a4) gene show features of lethal restrictive dermopathy. J Cell Biol *161*, 1105-1115.

Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., St Onge, R.P., Tyers, M., Koller, D., *et al.* (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science *320*, 362-365.

Holland, P.W.H., Garciafernandez, J., Williams, N.A., and Sidow, A. (1994). Gene Duplications and the Origins of Vertebrate Development. Development, 125-133.

Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and Predicting Haploinsufficiency in the Human Genome. PLoS genetics *6*.

Huguet, G., Ey, E., and Bourgeron, T. (2013). The genetic landscapes of autism spectrum disorders. Annual review of genomics and human genetics *14*, 191-213.

Hwang, Y.C., Lin, C.C., Chang, J.Y., Mori, H., Juan, H.F., and Huang, H.C. (2009). Predicting essential genes based on network and sequence analysis. Molecular bioSystems *5*, 1672-1678.

International Schizophrenia, C., Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748-752.

Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y.H., Yamrom, B., and Wigler, M. (2015). Low load for disruptive mutations in autism genes and their biased transmission. Proceedings of the National Academy of Sciences of the United States of America *112*, E5600-E5607.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., *et al.* (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216-221.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., *et al.* (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285-299.

Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. Genome research *20*, 1469-1481.

Jacquemont, S., Coe, B.P., Hersch, M., Duyzend, M.H., Krumm, N., Bergmann, S., Beckmann, J.S., Rosenfeld, J.A., and Eichler, E.E. (2014). A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. American journal of human genetics *94*, 415-425.

Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41-42.

Jeste, S.S., and Geschwind, D.H. (2014). Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nature reviews Neurology *10*, 74-81.

Ji, X., Kember, R.L., Brown, C.D., and Bucan, M. (2016). Increased burden of deleterious variants in essential genes in autism spectrum disorder. Proceedings of the National Academy of Sciences of the United States of America *113*, 15054-15059.

Johnsen, R.C., and Baillie, D.L. (1991). Genetic-Analysis of a Major Segment [Lgv(Left)] of the Genome of Caenorhabditis-Elegans. Genetics *129*, 735-752.

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome research *12*, 962-968.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., *et al.* (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231-237.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.

Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D., *et al.* (2012). Common genetic variants, acting additively, are a major source of risk for autism. Molecular autism *3*, 9.

Kohler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Ayme, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., *et al.* (2017). The Human Phenotype Ontology in 2017. Nucleic acids research *45*, D865-D876.

Koonin, E.V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature reviews Microbiology *1*, 127-136.

Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T.F., Morgan, H., Atienza-Herrero, J., Blake, A., Chen, C.K., Easty, R., Di Fenza, A., *et al.* (2014). The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic acids research *42*, D802-809.

Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., *et al.* (2015). Excess of rare, inherited truncating mutations in autism. Nat Genet *47*, 582-588.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics *9*.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. American journal of human genetics *95*, 5-23.

Leitner, Y. (2014). The co-occurrence of autism and attention deficit hyperactivity disorder in children - what do we know? Frontiers in human neuroscience *8*.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Lenz, L.S., Marx, J., Chamulitrat, W., Kaiser, I., Grone, H.J., Liebisch, G., Schmitz, G., Elsing, C., Straub, B.K., Fullekrug, J., *et al.* (2011). Adipocyte-specific Inactivation of Acyl-CoA Synthetase Fatty Acid Transport Protein 4 (Fatp4) in Mice Causes Adipose Hypertrophy and Alterations in Metabolism of Complex Lipids under High Fat Diet. Journal of Biological Chemistry *286*, 35578-35587.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., *et al.* (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, 886-897.

Li, B.S., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. Am J Hum Genet *83*, 311-321.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C., and Anckarsater, H. (2010). The Genetics of Autism Spectrum Disorders and Related Neuropsychiatric Disorders in Childhood. Am J Psychiat *167*, 1357-1363.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., *et al.* (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289-293.

Loh, P.R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., *et al.* (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet *47*, 1385-+.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord *30*, 205-223.

Lord, C., Rutter, M., and Lecouteur, A. (1994). Autism Diagnostic Interview-Revised - a Revised Version of a Diagnostic Interview for Caregivers of Individuals with Possible Pervasive Developmental Disorders. J Autism Dev Disord *24*, 659-685.

Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X.P., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A., *et al.* (2008). Highly parallel identification of essential genes in cancer cells. Proceedings of the National Academy of Sciences of the United States of America *105*, 20380-20385.

Luo, H., Gao, F., and Lin, Y. (2015). Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Scientific reports *5*.

Ma, D.Q., Salyakina, D., Jaworski, J.M., Konidari, I., Whitehead, P.L., Andersen, A.N., Hoffman, J.D., Slifer, S.H., Hedges, D.J., Cukier, H.N., *et al.* (2009). A Genome-wide Association Study of Autism Reveals a Common Novel Risk Locus at 5p14.1. Annals of human genetics *73*, 263-273.

Madsen, B.E., and Browning, S.R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS genetics *5*.

Malfatti, E., Lehtokari, V.L., Bohm, J., De Winter, J.M., Schaffer, U., Estournet, B., Quijano-Roy, S., Monges, S., Lubieniecki, F., Bellance, R., *et al.* (2014). Muscle histopathology in nebulin-related nemaline myopathy: ultrastrastructural findings correlated to disease severity and genotype. Acta neuropathologica communications *2*, 44.

Malhotra, D., and Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell *148*, 1223-1241.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., *et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet *82*, 477-488.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews Genetics *9*, 356-369.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.

Michalk, A., Stricker, S., Becker, J., Rupps, R., Pantzar, T., Miertus, J., Botta, G., Naretto, V.G., Janetzki, C., Yaqoob, N., *et al.* (2008). Acetylcholine receptor pathway mutations explain various fetal akinesia deformation sequence disorders. Am J Hum Genet *82*, 464-476.

Miklos, G.L.G., and Rubin, G.M. (1996). The role of the genome project in determining gene function: Insights from model organisms. Cell *86*, 521-529.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome biology *9*.

Mushegian, A.R., and Koonin, E.V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proceedings of the National Academy of Sciences of the United States of America *93*, 10268-10273.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., *et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.

NHLBI Exome Sequencing Project (ESP) Exome Variant Server [Internet]. Available from: http://evs.gs.washington.edu/EVS/.

Nijman, S.M. (2011). Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS letters *585*, 1-6.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., *et al.* (2012a). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science *338*, 1619-1622.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., *et al.* (2012b). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246-250.

Ozonoff, S., Heung, K., Byrd, R., Hansen, R., and Hertz-Picciotto, I. (2008). The onset of autism: patterns of symptom emergence in the first years of life. Autism research : official journal of the International Society for Autism Research *1*, 320-328.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. Cell *155*, 1008-1021.

Park, D., Park, J., Park, S.G., Park, T., and Choi, S.S. (2008). Analysis of human disease genes in the context of gene essentiality. Genomics *92*, 414-418.

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS genetics *9*, e1003709.

Piazza, R., Valletta, S., Winkelmann, N., Redaelli, S., Spinelli, R., Pirola, A., Antolini, L., Mologni, L., Donadoni, C., Papaemmanuil, E., *et al.* (2013). Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. Nature genetics *45*, 18-24.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., *et al.* (2010a). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368-372.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., *et al.* (2010b). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368-372.

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., *et al.* (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185-190.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., *et al.* (2015). ClinGen - The Clinical Genome Resource. New Engl J Med *372*, 2235-2242.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847-856.

Robinson, E.B., St Pourcain, B., Anttila, V., Kosmicki, J.A., Bulik-Sullivan, B., Grove, J., Maller, J., Samocha, K.E., Sanders, S.J., Ripke, S., *et al.* (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nat Genet *48*, 552-+.

Ronald, A., and Hoekstra, R.A. (2011). Autism Spectrum Disorders and Autistic Traits: A Decade of New Twin Studies. Am J Med Genet B *156B*, 255-274.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., *et al.* (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet *46*, 944-+.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., *et al.* (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863-885.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., *et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237-241.

Sanders, S.J., Xin, H., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Shan, D., *et al.* (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215-1233.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C.M., and Reichenberg, A. (2014). The Familial Risk of Autism. Jama-J Am Med Assoc *311*, 1770-1777.

Schinzel, A., and Giedion, A. (1978). A syndrome of severe midface retraction, multiple skull anomalies, clubfeet, and cardiac and renal malformations in sibs. American journal of medical genetics *1*, 361-375.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev *19*, 212-219.

Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome biology *6*, R33.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome research *13*, 2498-2504.

Shim, J., Moulson, C.L., Newberry, E.P., Lin, M.H., Xie, Y., Kennedy, S.M., Miner, J.H., and Davidson, N.O. (2009). Fatty acid transport protein 4 is dispensable for intestinal lipid absorption in mice. Journal of lipid research *50*, 491-500.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. Science *319*, 617-620.

Skarnes, W.C., Rosen, B., West, A.P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A.O., Thomas, M., Harrow, J., Cox, T., *et al.* (2011). A conditional knockout resource for the genome-wide study of mouse gene function. Nature *474*, 337-342.

Smoller, J.W., Craddock, N., Kendler, K., Lee, P.H., Neale, B.M., Nurnberger, J.I., Ripke, S., Santangelo, S., Sullivan, P.F., and Consortium, P.G. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet *381*, 1371-1379.

Stangenberg, M., Lingman, G., Roberts, G., and Ozand, P. (1992). Mucopolysaccharidosis-Vii as Cause of Fetal Hydrops in Early-Pregnancy. American journal of medical genetics *44*, 142-144.

State, M.W., and Levitt, P. (2011). The conundrums of understanding genetic risks for autism spectrum disorders. Nature neuroscience *14*, 1499-1506.

State, M.W., and Sestan, N. (2012). Neuroscience. The emerging biology of autism spectrum disorders. Science *337*, 1301-1303.

Steinberg, J., Honti, F., Meader, S., and Webber, C. (2015). Haploinsufficiency predictions without study bias. Nucleic acids research *43*.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for

clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics *133*, 1-9.

Stoner, R., Chow, M.L., Boyle, M.P., Sunkin, S.M., Mouton, P.R., Roy, S., Wynshaw-Boris, A., Colamarino, S.A., Lein, E.S., and Courchesne, E. (2014). Patches of Disorganization in the Neocortex of Children with Autism. New Engl J Med *370*, 1209-1219.

Szatmari, P., Paterson, A.D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X.Q., Vincent, J.B., Skaug, J.L., Thompson, A.P., Senman, L., *et al.* (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet *39*, 319-328.

The GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648-660.

Tonnsen, B.L., Boan, A.D., Bradley, C.C., Charles, J., Cohen, A., and Carpenter, L.A. (2016). Prevalence of Autism Spectrum Disorders Among Children With Intellectual Disability. Ajidd-Am J Intellect *121*, 487-500.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M., *et al.* (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature *459*, 528-533.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096-1101.

Weiner, D.J., Wigdor, E.M., Ripke, S., Walters, R.K., Kosmicki, J.A., Grove, J., Samocha, K.E., Goldstein, J., Okbay, A., Bybjerg-Gauholm, J., *et al.* (2016). Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Bioarxiv.

Weiss, L.A., Arking, D.E., and Consortium, J.H.A. (2009). A genome-wide linkage and association scan reveals novel loci for autism. Nature *461*, 802-U862.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research *42*, D1001-D1006.

White, J.K., Gerdin, A.K., Karp, N.A., Ryder, E., Buljan, M., Bussell, J.N., Salisbury, J., Clare, S., Ingham, N.J., Podrini, C., *et al.* (2013). Genome-wide generation and

systematic phenotyping of knockout mice reveals new roles for many genes. Cell *154*, 452-464.

Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., *et al.* (2013a). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell *155*, 997-1007.

Willsey, A.J., Sanders, S.J., Li, M.F., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., *et al.* (2013b). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. Cell *155*, 997-1007.

Willsey, A.J., and State, M.W. (2015). Autism spectrum disorders: from genes to neurobiology. Current opinion in neurobiology *30*, 92-99.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., *et al.* (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901-906.

Wray, N.R., Lee, S.H., Mehta, D., Vinkhuyzen, A.A., Dudbridge, F., and Middeldorp, C.M. (2014). Research review: Polygenic methods and their application to psychiatric traits. Journal of child psychology and psychiatry, and allied disciplines *55*, 1068-1087.

Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS genetics *6*, e1000864.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet *89*, 82-93.

Xu, X.X., Wells, A.B., O'Brien, D.R., Nehorai, A., and Dougherty, J.D. (2014). Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. Journal of Neuroscience *34*, 1420-1431.

Yonan, A.L., Alarcon, M., Cheng, R., Magnusson, P.K., Spence, S.J., Palmer, A.A., Grunn, A., Juo, S.H., Terwilliger, J.D., Liu, J., *et al.* (2003). A genomewide screen of 345 families for autism-susceptibility loci. American journal of human genetics *73*, 886-897.

Zhan, T.Z., and Boutros, M. (2016). Towards a compendium of essential genes - From model organisms to synthetic lethality in cancer cells. Crit Rev Biochem Mol *51*, 74-85.

Zhang, F., and Lupski, J.R. (2015). Non-coding genetic variants in human disease. Human molecular genetics *24*, R102-R110.

Zhang, M., Zhu, C., Jacomy, A., Lu, L.J., and Jegga, A.G. (2011). The orphan disease networks. American journal of human genetics *88*, 755-766.